Data Sharing Toolkit

Module 4

Anonymising data in agriculture

Guide

Anonymising data in agriculture

Purpose of this guide

One of the biggest barriers to sharing data are legitimate concerns over breaking legal obligations, security breaches or causing harm to individuals, communities or society. This guide will help grantees to navigate these concerns.

Specifically it will help explain the techniques for anonymising personal data in agricultural development projects and includes:

- Definition of personal data
- Role of data protection regulations
- Considerations of negative and positive impacts from data
- Practical mitigating actions

This guide is a subset of a larger guide in the Data Sharing Toolkit on managing risks to minimise harmful impacts when sharing data. The larger guide includes information on commercial and personal data, rights and permissions to use data, and encouraging best practice.

This document is not legal advice and if you are uncertain you should seek guidance from a legal professional.

When to use this guide

Concept | Proposal | Agreement | **Active**

- At the beginning of a project involving collection, access, use or sharing of data about people.
- When grantees are unsure whether they can share data containing personal information.
- When grantees need to maximise the utility of data while protecting the rights of individuals.
- When grantees need to assess what data they need to collect and keep in order to achieve the project objectives.

Quick links

Introduction

When designing a project with a long-term goal, it's important to balance collecting the data we need with minimising the data management risks. Some of this important data may be personal or societal data that cannot simply be published 'as is' due to data protection regulations and privacy concerns. However, with the application of appropriate techniques and processes, data can be made 'as FAIR and open as possible'[1] in a way that adheres to data protection regulations, protects privacy and avoids harm.

This guide looks at how anonymisation techniques can be applied to reduce the risks of re-identification and possible harm resulting from sharing data about people.

This guide draws on a wide body of evidence and best practice including guidance from the UK Information Commissioner's Office (ICO), the European Data Protection Board (EDPB) and the UK Statistics Authority. This guide compiles several standards followed by the industry and draws from current UK legislative mandates widely recognised as best practice.

If at any point you are concerned about potential harm we would encourage talking to data or legal experts.

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

Working through a practical example

In this guide we will work through a practical example of how to apply a number of techniques to anonymise data. For this we have made up a synthetic dataset that represents the types of information collected by a project collating soil health information. Typical goals for this type of project often include supporting small farms with tailored advice on improving yields, and providing aggregated data to national or international monitoring programmes.

Name	Gender	Date of birth	Farm location	Main crop	2019 yield	Soil density	Soil ph	Spouse name	Spouse profession
Amit Rao	M	10/07/1948	20.827580, 80.381560	Rice	100	1.5	6.8		
Priya Sethi	F	04/01/1982	20.770984, 80.428165	Rice	97	1.4	6.8	Raj Sethi	Farmer
Sam Bhalsod	M	01/07/2000	20.709031, 80.474158	Wheat	94	1.3	6.8		
Anjali Patnaik	F	25/12/1978	20.624257, 80.538830	Wheat	91	1.2	6.8	Amit Patnaik	Mechanic
Rakesh Batra	M	21/06/1982	20.621526, 80.568366	Wheat	88	1.5	6.8	Neha Batra	Midwife
Sanjana Mistry	F	17/12/1965	20.593900, 80.573586	Coffee	85	1.4	6.8	Soham Mistry	Teacher
Neeraj Lal	M	12/06/1949	13.469266, 77.168009	Coffee	82	1.3	6.8		
Sakshi Mohindra	F	07/12/1991	13.463363, 77.181036	Coffee	79	1.2	6.8	Vishal Mohindra	Doctor
Pranav Chaudhary	M	03/06/1996	13.454146, 77.184920	Maize	76	1.5	6.8	Riya Chaudhary	Farmer
Aswini Doshi	F	28/11/1962	13.438674, 77.172517	Wheat	73	1.4	6.8	Arjun Doshi	Mayor

Step 1: Establish the lawful and ethical basis

Data protection regulations across the world are designed to enable personal data to be 'processed', meaning collecting, accessing, using and sharing data, while minimising the risk of harmful impacts.

These regulations typically outline three key elements:

1. The lawful basis for using and sharing personal data
2. The rights of the person the data is about (the data subject)
3. Liabilities and penalties

Personal data is defined by the United Nations as 'information relating to an identified or identifiable natural person'.[2]

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 figure below provides some examples of personal data and other data about people.

Once personal data is anonymised and the risk of re-identification is sufficiently small, it is no longer subject to data protection regulations, however there may still be ethical implications.

[2] United Nations (2018), 'Personal Data Protection and Privacy Principles', https://archives.un.org/sites/archives.un.org/files/_un-principles-on-personal-data-protection-privacy-hlcm-2018.pdf

At this stage it is also important to consider the wider implications of collecting, using and sharing data. Tools like the ODI Data Ethics Canvas can help you consider important questions such as the limitations and biases in the data and whether it might adversely affect particular demographics or groups of people.

Staying safe

If you are concerned about the lawful basis for processing personal data, or unsure how to interpret data protection regulations, we encourage you to contact a legal professional.

As a minimum we strongly encourage organisations to communicate openly about what kind of data they hold, even if the data itself cannot be made openly available. Being open about data held can build trust with those the data is about, and those using it.[3]

Applying Step 1 to our example dataset

As our synthetic dataset is made up of fictional people, however, when dealing with a real data set of its type, we would note that the personal data includes the name and date of birth, and gender, which is also sensitive data. National and global legislation relating to using and sharing personal and sensitive data would need to be reviewed. It is likely that personal data would be automatically removed or suppressed to avoid any legal or ethical problems.

[3] Open Data Institute (2016), 'Openness principles for organisations handling personal data', https://theodi.org/article/openness-principles-for-organisations-handling-personal-data/, Accessed August 2020

What about farm location?

Farm location is not personal data by the UN definition, as it does not relate to a person, it relates to a farm. It is important not to get personal and non-personal data confused as the legislation will only apply to personal data, meaning that you can retain farm location without a lawful basis as stipulated by data protection regulations. However, there are legal and ethical implications to consider:

- If the precise location of the farm is not necessary for the project's goals there is no lawful basis for storing and sharing it.
- There might be a chance in some areas that combining precise farm location data with the farmers' gender, age or number or marital status could lead to farms managed or owned by women, elderly people or people living alone being targeted for aggressive sales techniques, sabotage or theft.

For this reason we might classify farm location as sensitive data, meaning data that is not already public and might cause harm if disclosed.

Step 2: Set objectives

Anonymisation should be done in a way that maintains as much of the intended utility of the data, while also protecting privacy.

In order to do this it is a good idea to outline some potential ways in which data might be used. In our example, our dataset is being used as part of a project to better understand soil health in a region. But it would also be desirable to use the data to assess diversity and inclusion in our project's implementation.

Applying Step 2 to our dataset

We want to make this data openly available because:

- We want to contribute our data to a predictive model about soil health and crop yields.
- We want people to be able to see which demographics, ages and locations are involved in and benefit from the project.

Step 3: Assess risk

It is important to understand that data protection does not necessarily require anonymisation to be completely risk free – but you must be able to mitigate the risk of re-identification until it is sufficiently small.

To do this, identify characteristics in your dataset that can directly or indirectly be used to identify a person: the **personal** data.[4]

Next, determine which of these characteristics carry a potential threat to an individual, then assess whether they pose a 'normal' or 'high' risk of re-identification. Risk management can be a complex topic and depends heavily on context and domain knowledge, but as a general rule a risk is classified as 'high' if both its likelihood and its impact are high, or if the likelihood is lower but the impact would be unacceptable. See our guide on managing risk for more detail.

It might be helpful to think about risk based on the following:

1. Probability of an attacker attempting to re-identify an individual
2. Probability of an attacker in succeeding to re-identify an individual
3. Consequences to the individual who has been identified

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

Applying Step 3 to our example dataset

We identified the personal data in Step 1 as part of the legal and ethical considerations:

- Names (of both the farmer and their spouse, if they have one)
- Gender (which is also sensitive data)
- Date of birth

We also identified that farm location is not strictly personal data, however, there may still may be risks related to disclosing a location independently or in combination with other personal data.

Example

- **Concern**: If precise farm locations are included in the data along with age ranges and marital status, criminals could potentially identify farms where older people live alone and target those individuals for theft or fraud.
- **Risk score**: This is made up of two values, the **likelihood score** and the **impact score**. The likelihood might be assessed as medium (depending on the history of similar crimes in the area, so local context is crucial) but the **impact** would be severe if it happened. So the risk score would be ‘high’.
- **Mitigation**: an appropriate mitigating action would be to suppress the precise farm location in the data.

Minimising risk

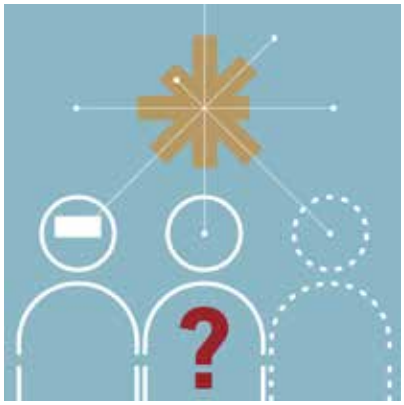
Your objective (and ours) is to safely publish FAIR and safeguarded data.

Any data anonymisation has to consider a balance between risk and utility – how much do you minimise risk while trying to keep the data as useful as possible. One of the challenges with open data is that it is impossible to predict all of the ways in which the data could be used by others.

If you are unsure, we encourage you to engage with experts in data collection and management, such as the relevant national organisation specialising in information governance and data protection.

Where open publication would be completely inappropriate or potentially risky it is still possible to share the data under more controlled conditions, for example through a data sharing agreement with a limited set or group of individuals. The Agricultural Data Spectrum in Module 5 may help you to consider who needs to have access to the data, and our guides on deciding how to provide access to data and designing data sharing agreements in Module 5 will help you become familiar with the methods and tools available.

Step 4: Anonymise personal data



There are several techniques that can be used to anonymise data. In the majority of cases, several techniques may need to be applied to lower the risks of re-identification.

Choosing the right combination of techniques will depend on both the intended use and the level of risk related to each characteristic in the dataset. A record should be kept on how each technique is applied and the reasons for the choice.

The EU Data Protection Working Party has produced a useful report outlining anonymisation techniques, robustness and typical mistakes.[5] We have summarised these techniques below.

[5] Data Protection Working Party (2014), Article 29, Opinion 05/2014 on Anonymisation Techniques, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, Accessed August 2020

Anonymisation techniques

Suppression

Suppression simply involves removing data from the dataset, such as any identifier or person's name. Suppression is best applied on any direct identifiers. It is important to carefully consider the use cases from Step 1 before simply deleting data, otherwise, the overuse of suppression is highly likely to reduce the utility of data.

Randomisation

Randomisation is a family of techniques that alters the data by adding and subtracting small amounts to the numbers in columns of numeric values, or shuffling the order of values in numeric or text columns, all while maintaining the characteristics and patterns in the dataset as a whole. This adds uncertainty to the data in order to remove the strong link to the individual.

Randomisation by itself does not mask individuals in the dataset, it does however make it challenging to make precise inferences about those individuals. Additional techniques may be required to ensure that records cannot be used to identify a single individual.

Generalisation

This approach dilutes the attributes of a dataset by making them more general: a dataset that records cities might be aggregated into regions; a dataset that records weeks, into months. Generalising modifies the respective scale or order of magnitude of a data set, thereby making it less specific, and harder to use for identifying individuals.

Whilst generalisation can be effective to prevent re-identification, it does not allow effective anonymisation in all cases. It may still be possible for an attacker to infer information about an individual using some other combination of attributes to link two or more records to a specific individual or group, or single out an individual by using a combination of filters until only one record is left.

Pseudonymisation

Pseudonymisation is a useful security measure but not a method of anonymisation. It consists of replacing a direct identifier like a name or ID number with an artificial identifier or pseudonym. It has been included here because it can reduce direct re-identification, however indirect identification is still possible given enough additional data.

Applying Step 4 to our example dataset

If we were to suppress all the personal data, then this dataset might not be useful to fulfil our objectives, so we need to consider other methods.

It is useful to consider what needs to be kept to fulfil our objectives:

Personal data field	Objective: Contribute to a predictive model about soil health and crop yields	Objective: See which demographics benefit from the project
Name	No	No
Gender	No	Yes
Date of birth	No	Yes
Spouse name	No	No

From this it is clear we can **suppress** the names of individuals in this dataset.

We can also **aggregate** the date of birth into an age range, for example. 0–18,19–69,70–120, or perhaps just retain the year. The choice will very much depend on how flexible we want to be with the second objective.

Original (Date of birth)	Anonymised (age band)
10/07/1948	70–120 years
04/01/1982	18-45 years

At this point, while the people are not directly identifiable, a combination of the other characteristics recorded in the dataset may allow people to be re-identified.

The next most identifiable piece of information is probably the farm location, which is currently given as exact latitude and longitude coordinates. Given its specific nature, this data can be linked to a particular farm where only a limited number of people are likely to live or work. We know that farm location is not personal data, but if there are concerns about the impacts on people from this data, it could also be **aggregated**:

Original (Farm location)	Anonymised (Region)
20.593900, 80.573586	Maharashtra
13.469266, 77.168009	Karnataka

Despite these steps, it may still be possible to identify someone from the remaining data. For example, there may be combinations of categories within the dataset that yield only one possible result. We need to test the resilience of our anonymisation.

Step 5: Testing resilience

Anonymisation is considered successful when the risk of re-identification is low. To test this you should carry out resilience tests.

This step will also help you analyse the balance between risk and utility. As before, if you are worried, take a more cautious approach and seek advice from experts.

Testing resilience can be done by asking three questions:

1. Is it possible to single out an individual by filtering the data set using one or more attributes to yield a single result?
2. Is it possible to link records relating to an individual using data already available in other data sets?
3. Can information be inferred concerning an individual within the same dataset?

You should also consider viewing the data from the perspective of someone who may have an undesirable interest in the dataset. You may have identified these groups or individuals during the risk assessment when considering the intentions of possible attackers. If you cannot think of any, but you labelled data ‘high risk’, it could be that the risk is not as high risk as you first thought. Defining levels of risk can be subjective, but there are techniques you can use to help with this. See the guide on managing risk in Module 4 for more information.

If you answered ‘yes’ to any of the above, go back to Step 3.

Applying Step 5 to our example dataset

Let's apply the three questions to the current in progress dataset:

Attribute	Example value	Q1: Is it possible to single out an individual?	Q2: Is it possible to link records relating to an individual using data already available to people?	Q3: Can information be inferred concerning an individual within the same dataset?
Name	[ALREADY REMOVED]			
Gender	F	N	N	N
Age band	40-64	N	N	N
Region	Karnataka	N	N	N
Main crop	Maize	N	N	N
2019 yield	73	N	N	N
Soil density	1.4	N	N	N
Soil ph	6.8	N	N	N
Spouse name	[ALREADY REMOVED]			
Spouse profession	Mayor	Y	Y	N

For a small dataset like our example, Question 1 can be quick to check. In a larger dataset with hundreds or even thousands of rows, techniques like filters and pivot tables can help to identify whether there are any unique values or very small groups of attributes that might single out an individual.

Local and contextual knowledge will also be important here. Even without filtering we could infer that certain professions such as Mayor will be unique to an individual and a region, so we need to look for a way to anonymise this.

Looking back at typical objectives for this type of project, ‘Profession’ may not be strictly required to fulfill our objectives however it might be a useful proxy for education level or family income. There may be anonymisation techniques we can apply to keep it.

One potential technique is K-Anonymisation, which simply means changing the value in the data, so it is no longer unique but keeps the same or similar meaning. Here we change the role to be the same as a number of other people (either in the dataset or in real life). Perhaps a logical change would be ‘Mayor’ to ‘Politician’? This would be less exact but still not technically wrong.

This gives us the final dataset:

Gender	Age band	Region	Main crop	2019 yield	Soil density	Soil ph	Spouse profession
M	65-90	Maharashtra	Wheat	95	1.5	6.8	
F	40-64	Maharashtra	Wheat	97	1.4	6.8	Farmer
M	18-39	Maharashtra	Wheat	94	1.3	6.8	
F	40-64	Maharashtra	Wheat	91	1.2	6.8	Mechanic
M	40-64	Maharashtra	Wheat	88	1.5	6.8	Midwife
F	40-64	Maharashtra	Coffee	85	1.4	6.8	Teacher
M	65-90	Karnataka	Coffee	82	1.3	6.8	
F	18-39	Karnataka	Coffee	79	1.2	6.8	Doctor
M	18-39	Karnataka	Maize	76	1.5	6.8	Farmer
F	40-64	Karnataka	Maize	73	1.4	6.8	Politician

Step 6: Write a plan to minimise risk of re-identification

You should create a plan to monitor the risk of re-identification and handle any potential impacts. Even when you are satisfied the risks are acceptably small, if they do arise you may need to take action to protect individuals or manage reputational or financial impacts.
For more detail you can refer to the guides Managing Risk with Personal Data in Module 6 which covers how to identify risks, and Developing a Data Management Plan in Module 7 which includes a sample privacy impact assessment.

A clear plan prepared ahead of time will help you to respond appropriately. Include roles and responsibilities that describe who does what in a given situation to help clarify the actions that need to be put in place in case of re-identification.

This plan should be shared with a legal professional or other authority for them to review and decide if the methods and procedures you applied pose minimal risk to re-identification.

Applying Step 6 to our example dataset

An outline plan might look something like this:

Data protection

All participants are required to complete a questionnaire either online or in person, and their consent to collect, store and process their personal data has been obtained and recorded in our system. If a participant revokes that consent, their data will be deleted from the research dataset, however a record of their initial participation will be retained to ensure we keep their preferences on file and can avoid contacting them again. A full privacy impact assessment for the project has been conducted and signed off by the Head of Information Governance, and any concerns should be raised with them as soon as possible.

Data sharing

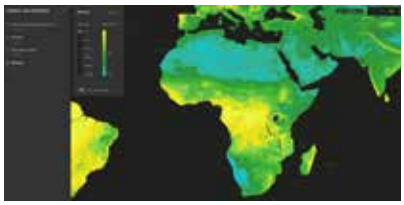
For the duration of the funded project, the Analytics team will be responsible for data capture and management. Any queries about protocols, storage and access should be emailed to the Analytics Team Lead.

Data management

After the project is finished, participant data will be made available to recognised research organisations via our secure portal. Access will be contingent on completion of a data sharing agreement.

Anonymised and aggregated versions of the data will be published under an open license, managed and maintained by the Information Governance team. For more information and to report any issues, contact the Head of Information Governance.

Step 7: Publish the data, anonymisation details and risk assessment



Once the dataset is anonymised and the risk of re-identification is low, it can be published for others to use.

In addition to the data, it is highly recommended that the following are also published:

- The objectives for your anonymisation
- The techniques applied to anonymise the dataset, for example 'Low numbers have been suppressed and replaced by 0, and sampling locations have been aggregated by region.'.
- Your re-identification risk assessment
- Your plan to mitigate risks
- Details of how people can reach out with questions and concerns

Applying Stage 7 to our example dataset

If we've taken all the necessary steps to protect our data subjects while preserving the utility of the data, we might expect to see this type of data used in things like the ISRIC SoilGrid[6] soil health mapping project.

[6] ISRIC SoilGrid (2020), https://www.isric.org/explore/soilgrids, Accessed July 2020

Additional online resources

- Responsible Data Guidelines, CGIAR. Managing privacy and personally identifiable information in research project data lifecycle; a resource for researchers in agricultural projects.
- Anonymisation: managing data protection risk code of practice, Information Commissioner's Office UK (ICO). Published by the ICO in 2015, provides practical advice on methods for anonymising data and the associated risks.
- Guide on intruder testing, Office for National Statistics UK (ONS). This guide was created by the ONS to show the steps organisations – but mainly governmental departments – need to follow to ensure they are meeting ethical and legal requirements in protecting individuals, households and businesses. The guide demonstrates the steps related to conducting an intruder testing that assess the likelihood of someone to be identified in a dataset that is published in the open domain.

- Policy for social survey microdata, ONS. This guide was created by the ONS to show the steps required prior to publishing data on the public domain to ensure that people's personal and sensitive data is protected from harm. Although the guide does not use the term anonymisation – it uses 'disclosure control' – the guide can easily be compared to the steps explained in *Anonymising data in times of crisis*.
- Opinion 05/2014 on Anonymisation Techniques, EUData Protection Working Party. Explores common anonymisation techniques, risks and common mistakes when applying each technique.

- Anonymisation: register of actors, ODI/Eticas.
  A list of actors in the field, from academia to the private sector, who can help with anonymisation.
- Anonymisation: A short guide, ODI/Eticas.
  A short guide on practical anonymisation from Eticas Research and Consulting to help inform and steer its work, and provide a reference for anyone interested in the topic.
- Anonymisation: case studies, ODI/Eticas.
  Examples of anonymisation for Health data, geolocation data and statistics.

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.1116758.1

cabi.org | theodi.org | gatesfoundation.org

CABI Data Sharing Toolkit

open data institute

BILL & MELINDA GATES foundation