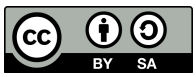
OPEN FOR FEEDBACK

Anonymising data in times of crisis

Open Data Institute

About

This guide has been produced by the Open Data Institute, and published in May 2020. Its lead authors are David Tarrant and Violeta Mezeklieva with contributions from Olivier Thereaux, Renate Sampson and Jeni Tennison.



How can it be improved? We welcome suggestions from the community in the comments.

Introduction

In times of crisis, we need as much data as possible to be as open as possible to help people make the right decisions. 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 open as possible’ 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 existing guidelines including from the Information Commissioner's Office (ICO), the National Health Service (NHS UK), European Data Protection Board (EDPB) as well as the UK Statistics Authority. This guide compiles several standards followed by the industry and includes current legislative and regulatory mandates.

If at any point you are concerned about potential harm we would encourage talking to experts. As a minimum follow Step 1 to establish your ability to at least tell people that you have the data.

Working through a practical example

In this guide we work through a practical example. For this we have synthesised (made up) the following dataset:

Site	Call date	Name	Gender	Date of birth	Address	Postcode	Profession	Severity (1=HIGH)
111	18/03/2020	Zoe Bloggs	F	2005-06-03	12 Penn Street, Barking, London	IG11 0DI	Student	35
111	18/03/2020	Patricia Butcher	F	1949-11-10	27 Penelope Drive, Eastbourne	BN20 7UP	Retired	3
111	18/03/2020	Barry Titchmarsh	M	1981-10-08	19 Barber Close, Matlock, Derby	DE4 2ME	Gardener	90
999	18/03/2020	Gilderoy Lockhart	M	1946-02-28	27 Diagon Alley, Tower Hamlets, London	E1 0HP	Professor	1
999	14/04/2020	Arthur Smith	M	1922-03-15	155 Mount Park, Shewsbury	SY1 0AP	Retired	2
999	16/04/2020	Joe Quimby	F	1965-09-17	17 Prestige Place, Ringwood	BH24 1TM	Mayor	6

This data represents what might be typical of the type of data collected by a health service which is dealing with patients who potentially are infected with COVID-19. In this case the data is closely representative of the data collected when people in the UK telephone either the National Health Service helpline (111) or the emergency services (999).

STEP 1: Establish the lawful and ethical basis

Once personal data is anonymised, meaning the risk of re-identification is sufficiently small, it is no longer subject to data protection regulations (such as the General Data Protection Regulation (GDPR) or the UK Data Protection Act 2018).

However, prior to anonymising personal data you must assess the lawful and ethical basis for using it and if necessary undertake a Privacy Impact Assessment[1] - particularly if the data you are using is special category data – that is, sensitive data.

The law does not prevent collecting, using or sharing of personal data where it is necessary and proportionate. One of the most important aspects of data protection law is the requirement to establish a lawful basis for processing personal data. The ICO has a useful guide to the lawful basis under data protection regulations[2]. They have also published a guide to help organisations navigate personal data related to Covid-19[3].

In addition to the lawful basis, the ODI Data Ethics Canvas helps consider the wider implications of collecting, using and sharing data[4]. The Data Ethics Canvas helps you consider important questions such as the limitations and biases in the data and if it might adversely affect particular demographics or other groups of people.

Regardless of how personal data is processed, people must be informed how data about them is being collected, used and shared.[5]

Staying safe

If you are concerned about the lawful basis for processing personal data we encourage you to contact your organisation's data protection officer who may be able to clarify this.

As a minimum we strongly encourage organisations to communicate openly about what kind of data they hold, even if the data itself can not be made openly available.

Application to our example dataset

As this dataset is made up of fictional people, we do not need a lawful basis for using it. Although there is ‘personal data’ in the dataset, it is not related to any living individual.

[1] Information Commissioner’s Office (2019), Data protection impact assessments
[2] Information Commissioner’s Office (2020), Lawful basis for processing
[3] Information Commissioner’s Office (2020), Data protection and coronavirus information hub
[4] Open Data Institute (2019), The Data Ethics Canvas
[5] Information Commissioner’s Office (2019), Right to be informed

From an ethical viewpoint, there is a small risk that someone might believe that Gilderoy Lockhart (a half-blood professor from the Harry Potter series) does live in Tower Hamlets, or that Joe Quimby (the Mayor from the Simpsons) lives in Ringwood. This person then might go to that street address (if our fictional street even exists) or enter the postcode (which does exist) in a SatNav and end up annoying whoever may live in that location, but the risk of that is acceptably low. So we can continue.

STEP 2: Set objectives

Anonymisation should be done in a way that aims to maintain as much of the intended utility while also protecting privacy.

In order to do this it is a good idea to outline some potential use cases:

- People should be able to use the anonymised data to analyse the demand for NHS services (111/999). They should be able to do this over time and on a reasonably local basis.
- People should be able to use the anonymised data to make intelligent decisions on when to go grocery shopping.
- People should be able to use anonymised data to…

Staying safe

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

Any data anonymisation has to consider a risk/utility balance; that is how much do you change the dataset to minimise risk while trying to preserve the utility. One of the challenges with open data is that you will never know all of the ways in which the data could be used. If you are unsure, we encourage you to engage with experts or take a more cautious approach. Such approaches could include sharing data with a limited set or group of individuals prior to making the data open. The ODI's Data Spectrum helps illustrate the different types of access to data including shared and open data[6].

Application to our dataset

We want to make this data openly available because:

1. We want to allow local councillors and support groups to get a view on the potential spread of the virus geographically.
2. We want people to be able to see who (by demographics, age and location) needs the most support.

[6] Open Data Institute (2018), The Data Spectrum

STEP 3: Assess risk

It is important to understand that data protection does not require anonymisation to be completely risk free – you must be able to mitigate the risk of re-identification until it is sufficiently small[7].

To do this, identify characteristics in your dataset that can directly or indirectly be used to identify a person: the personal data[8].

Following this, you should determine which of these characteristics pose potential threat to harm an individual and label them as either posing a 'normal' or 'high' risk to re-identification.

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

a) Probability of an attacker attempting to re-identify an individual
b) Probability of an attacker in succeeding to re-identify an individual
c) Consequences to the individual who has been identified

For example, data about abortion is of interest to certain groups who may have unlawful or unethical intentions and who are therefore seeking to identify individuals. The environment and context for publishing or sharing this data has to be closely considered to protect these individuals from harm.

Application to our example dataset

Identify the personal data.

We definitely have:

- Name
- Gender (also sensitive data)
- Date of birth
- Address
- Postcode

[7] Information Commissioner's Office (2012), Anonymisation: managing data protection risk code of practice
[8] Information Commissioner's Office (2020), What is personal data?

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 its choice.

The EU Data Protection Working Party has produced a useful report outlining anonymisation techniques, robustness and typical mistakes[9]. We have summarised these techniques in this section.

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 noise or shuffling values 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 provide noise that makes precise inferences about those individuals challenging. Additional techniques may be required to ensure that a record cannot identify a single individual.

[9] Data Protection Working Party (2014), Article 29, Opinion 05/2014 on Anonymisation Techniques

Generalisation

This approach consists of generalising or diluting the attributes of data by modifying the respective scale or order of magnitude (that is, a region rather than a city, a month rather than a week).

Whilst generalisation can be effective to prevent re-identification, it does not allow effective anonymisation in all cases; in particular, it requires specific and sophisticated quantitative approaches to prevent linkability and inference.

Pseudonymisation

Pseudonymisation is a useful security measure but not a method of anonymisation. It consists of replacing a direct identifier with an artificial identifier or pseudonym. Pseudonymisation reduces direct re-identification however indirect identification is still possible given enough information.

Typical pseudonymisation techniques include encryption, hashing and tokenisation.

Application to our example dataset

If we were to suppress (remove) all the personal data, then this dataset would not be useful to fulfil any of our objectives. So we need to consider other methods.

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

Personal data field	Objective #1	Objective #2
Name	No	No
Gender	Potentially useful	Potentially useful
Date of birth	No	Yes
Address	Yes	Yes
Postcode	Yes	Yes

From this it is clear we can **suppress** the person's name.

At this point, while the person is not directly identifiable, a combination of the other characteristics will allow people to be re-identified easily, therefore we haven't finished yet.

The next most identifiable piece of information is probably the exact address, given that only a small number of people will live at a single address. We can **aggregate** this information to anonymise it:

Address (original)	Address (anonymised)
12 Penn Street, Barking, London	Barking, London
27 Penelope Drive, Eastbourne	Eastbourne
19 Barber Close, Matlock, Derby	Matlock, Derby
27 Diagon Alley, Tower Hamlets, London	Tower Hamlets, London
155 Mount Park, Shrewsbury	Shrewsbury

We could also do the same with the postcode, removing the last three digits to just identify regions in a city, for example, W1, W2 etc.

Aggregation can also be used to translate the date of birth into an age range, for example. 0–18,19–69,70–120.

Date of birth (original)	Age band (anonymised)
2005-06-03	0–18 years
1949-11-10	70–120 years
1981-10-08	19–69 years
1946-02-28	70–120 years
1922-03-15	70–120 years

At this point you may think you are done. However it may still be possible to identify someone from the remaining data. 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 risk/utility balance. As before, stay safe and if you are worried, take a more cautious approach.

This can be done by asking simple questions:

1. Is it possible to single out an individual?
2. Is it possible to link records relating to an individual using data already available?
3. Can information be inferred concerning an individual within the same dataset?

You should also consider putting yourself in the shoes of a stakeholder who may have an nefarious interest in the dataset. You may have identified these stakeholders during the risk assessment when considering intentions of attackers. If you cannot think of any stakeholders, but you labelled data 'high risk', it could be that the risk is not as high risk as you first thought.

If you answered *'yes'* to any of the above, go back to Step 3.

Application to our example dataset

Remember the questions:

1. Is it possible to single out an individual?
2. Is it possible to link records relating to an individual using data already available to people?
3. Can information be inferred concerning an individual within the same dataset?

Let's apply these to the dataset the current in progress dataset:

Feature	Example value	Q1	Q2	Q3
Site Type	111,999	N	N	N
Call date	2020-03-19	N	N	N
Name	Joe Bloggs	REMOVED FROM DATA		
Date of Birth	2005-06-03	REMOVED FROM DATA		
Age Band	0–18,19–69,70–120	N	N	N
Town/City	Matlock	N	Potentially in some sparsely populated areas. **Classifying as normal risk.**	N
Postcode area	IG11	N	Potentially in some sparsely populated areas. **Classifying as normal risk.**	N
Profession	Gerdenor, Mayor	Y	Y	Y

From this we can observe that profession, such as mayor, may be unique to an individual in an area.

Back to Step 2

'Profession' may not be required to allow our objectives to be fulfilled however there may be potential anonymisation techniques we can apply to keep it.

One potential technique is K-Anonymisation. Here we change the profession to be the same as at least N other people (either in the dataset or in real life). Perhaps a logical change would be 'Mayor' to 'Councillor'? This would be less exact but still not technically wrong.

This gives us the final dataset:

Site	Call Date	Gender	Age Band	Address	Postcode District	Profession	Severity (1=HIGH)
111	2020-03-18	F	0–18 years	Barking, London	IG11	Student	35
111	2020-03-18	F	70–120 years	Eastbourne	BN20	Retired	3
111	2020-03-18	M	19–69 years	Matlock, Derby	DE4	Gardener	90

999	2020-03-18	M	70–120 years	Tower Hamlets, London	E1	Professor	1
999	2020-04-14	M	70–120 years	Shrewsbury	SY1	Retired	2
999	2020-04-16	F	0–18 years	Ringwood	BH24	Councillor	6

STEP 6: Write a plan to mitigate risk

You should create a plan to monitor and handle any potential risks, for example when normal risks might require action.

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 data protection officer or other authority for them to review and decide if the methods and procedures you applied pose minimal risk to re-identification.

Application to our example dataset

Rather than apply and plan to mitigate risk regarding our dataset, we applied this step to the whole of this guide. The guide was reviewed internally by senior members of the ODI team prior to publication to ensure our advice is both accurate and set in the right tone that provides the right balance between risk and utility.

All of our guides are open for comment and if you have any questions or concerns please get in touch.

At time of writing, Olivier Thereaux (Head of Research and Development at the ODI) is the senior sponsor ultimately responsible for the content of this guide.

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

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

For help on publishing the data, take a look at this step-by-step guide to publishing data in times of crisis.

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

1. The objectives for your anonymisation
2. The techniques applied to anonymise the dataset
3. Your risk assessment
4. Your plan to mitigate risks
5. Details of how people can reach our with questions and concerns

Application to our example dataset

We are not going to do this with our synthesised dataset since we wanted to show the process that leads up to this step. The real dataset (anonymised) is available on the NHS website where you can read more about the objectives and limitations of the dataset.

Further resources

Anonymisation: managing data protection risk code of practice, ICO
Published by the ICO in 2015, provides practical advice on methods for anonymising data and the associated risks.

Guide on intruder testing, 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 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.