A Best Practice Approach to Anonymization

Elaine Mackey

Contents

Abstract

The need for clear guidance on anonymization is becoming increasingly pressing for the research community given the move toward open research data as common practice. Most research funders take the view that publicly funded research data are a public good which should be shared as widely as possible. Thus researchers are commonly required to detail data sharing intentions at the grant application stage. What this means in practice is that researchers need to understand the data they collect and hold and under what circumstances, if at all, they can share data; anonymization is a process critical to this, but it is complex and not well understood. This chapter provides an introduction to the topic of anonymization, defining key terminology and setting out perspectives on the assessment and management of reidentification risk and on the role of

E. Mackey (✉)
Centre for Epidemiology Versus Arthritis, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
e-mail: elaine.mackey@manchester.ac.uk

R. Iphofen (ed.), *Handbook of Research Ethics and Scientific Integrity*,
https://doi.org/10.1007/978-3-319-76040-7_14-1

anonymization in data protection. Next, the chapter outlines a principled and holistic approach to doing well-thought-out anonymization: the *Anonymisation Decision-making Framework* (ADF). The framework unifies the technical, legal, ethical, and policy aspects of anonymization.

Keywords
Anonymization · Anonymisation Decision-making Framework · Data environment · Personal data · General Data Protection Regulation

Introduction

Anonymization is essentially a process to render personal data nonpersonal. Given this description of anonymization, it may seem as if it is a simple process, but this is not the case. Anonymization is complex and not well understood, but nevertheless it is integral to the collection, management, and sharing of data appropriately and safely. The need for clear guidance on anonymization is becoming increasingly pressing, in particular for the research community given the move toward open research data as common practice. Most research funders take the view that publicly funded research data are a public good which should be made openly available, with as few restrictions as possible and, within a well-defined time period (see Concordat on Open Research Data 2016; Open Research Data Taskforce Report 2017, 2018). As a consequence researchers are frequently required to detail data sharing intentions at the grant application stage. What this means in practice is that researchers need to understand the data they collect and hold and under what circumstances, if at all, they can responsibly share data.

This chapter provides an introduction to the topic of anonymization; it is divided into two parts. In the first part, the key terminology associated with anonymization is defined within the framework of European (including UK) data protection legislation. The discussion then turns to consider how it is one thinks about the reidentification problem influences the way in which risk is understood and managed. The reidentification problem refers to the inherent risk of someone being reidentified in a confidentiality dataset. In the final section of part one, the role of anonymization in data protection is examined; there are two approaches, i.e., the absolute approach and the risk-based approach. These approaches take opposing views on what the risk of identification in a confidential dataset should realistically be.

In the second part of the chapter, a framework for doing well-thought-out anonymization is outlined, the *Anonymisation Decision-making Framework* (ADF), which is written up in a book of the same name (Elliot et al. 2016). The ADF is a first attempt to unify into a single architecture the technical aspects of doing anonymization with legal, social, and ethical considerations. The framework is underpinned by a relatively new way of thinking about the reidentification problem called the *data environment perspective*. This (data environment) perspective shifts the traditional focus on data toward a focus on the relationship between data and data

environment to understand and address the risk of reidentification (Mackey and Elliot 2013; Elliot and Mackey 2014).

What Is Anonymization

If you know anything about anonymization, you will invariably have noted the complex nature of the topic. The term, itself, is not uniformly described, nor is there complete agreement on its role in data protection, or on how it might be achieved. In this section, some of the inherent complexities underpinning anonymization are drawn out and current thinking on the topic outlined which can be seen as in line with the new General Data Protection Regulation (GDPR 2016/679) and the UK's Data Protection Act (DPA 2018).

The terms anonymization and de-identification are sometimes taken to mean the same thing. However these two terms have quite different meanings when applied in a European (and UK) context. As an aside, it is worth noting that the USA, Canada, and Australia use this terminology differently from Europe. Anonymization as a concept is understood within a legal framework, so to elaborate further, one must look to GDPR and in particular at how the Regulation defines personal data. GDPR defines personal data as meaning:

> *any information relating to an identified or identifiable natural persons'; **an identified natural person is one who can be identified, directly or indirectly*** . . . Article 4(1)

The key part of the definition of personal data of interest for the purpose of describing anonymization is the sentence highlighted in bold, namely, that an identified person is one that can be identified from data either:

1. Directly or
2. Indirectly

The process of de-identification refers to the removal, replacement, and/or masking of direct identifiers (sometimes called formal identifiers) such as name, address, date of birth, and unique (common) reference numbers and as such it addresses *no more* than the first condition, i.e., the risk of identification arising directly from data (Elliot et al. 2016).

The process of anonymization, in contrast, should address both conditions 1 and 2, i.e., the risk of identification arising directly and indirectly from data. The removal of direct identifiers is necessary, but rarely sufficient on its own for anonymization. Thus, one will be required to either further mask or alter the data in some way or control the environment in which the data exists (Elliot et al. 2016). Anonymization, therefore, might be best described as a process whereby personal data, or personal data and its environment, are modified in such a way as to render data subjects no longer identifiable. Data that has undergone the process of anonymization is considered anonymous information and is not in scope of GDPR. Recital 26 of GDPR

stipulates that "the principles of data protection should therefore not apply to anonymous information" (2016/679).

There is a third term critical to this discussion, that of "pseudonymization." It is a concept newly introduced into data protection legislation by GDPR. Pseudonymization is defined as meaning, "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject *without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures* to ensure that the personal data are not attributed to an identified or identifiable natural person" (GDPR Article 4(5) 2016/679). Reading this definition, in particular the text in bold, we might reasonably surmise that pseudonymization maps on to the description given previously of de-identification, most especially, that pseudonymization like de-identification addresses no more than the risk of identification arising directly from data. The idea that pseudonymization addresses only part of the reidentification risk fits with the very important point that, within the framework of GDPR, data that has undergone the process of pseudonymization *is considered personal data*.

The discussion has thus far provided a description and definition of the core concepts of anonymization, de-identification, and pseudonymization (for the remainder of the chapter, the term de-identification will not be used). The next step is to apply the concepts of anonymization and pseudonymization in practice, and this is where it gets complicated. How these concepts are applied will be shaped by how one thinks about and manages the reidentification problem. Traditionally, researchers and practitioners have addressed the reidentification problem by focusing (almost exclusively) on the data to be shared meaning that the models built to assess and control reidentification risk, whilst statistically sophicated, are largely based on *assumptions* about real world considerations such as the how and why of a reidentification attempt (Mackey and Elliot 2013). Furthermore, this approach does not take into account the issue of perspectives, that is who is looking at the data; a 'data controller' or 'data processor' or 'data user'. A data centric approach is likely to lead to the conclusion that data that has undergone the process of pseudonymization is *always* personal data, but there may be particular sets of circumstances under which one can argue that this is not the case, remember the defintion of anonymization just given that it may be achieved by modifying data or data and its environment. To explain, and this point is fundamental – data does not exist in a vacuum. Rather, data exists in an environment that Mackey and Elliot (2013) describe as having four additional components to the data to be shared, that of the presence or absence of other data (that can be potentially linked to the data to be shared), agents (which incorporations the notion of who is looking at the data), the presence or absence of governance processes and infrastructure (expanded on in section "Data Environment Perspective"). Thinking about data in relation to environments puts a spotlight on a crucial point that data should not be viewed as a static unchanging object; for example as 'anonymous' for ever once the researcher has removed/masked direct identifiers and masked/altered (some of) the statistical properties of the data. Rather data should be understood as a dynamic object, its status as personal data or anonymous information is dependent on the inter- relationship

between data and environment – not just data. If factors are changed with either or both data and environment the status of the data, as personal data or anonymous information, might well change. This point about the dynamic nature of data underpins Mourby et al. (2018) argument that data that has undergone the process of pseudonymization may or may not be personal data – it will crucially depend on the data - data environment relationship (also see Elliot et al. 2018). To illustrate this point further, imagine the following scenario: one detailed dataset and two possible share environments, one is a "open publication" environment the other a highly controlled "safe haven" environment. The dataset is described thus: direct identifiers are removed (i.e., participants' names, address, date of birth); it contains demographic data (3-digit postcode, age in yearly intervals, gender, and ethnicity), clinical data (current diseases and conditions, past medical history, and concomitant medication), and event date data (hospital admissions and outpatient clinic attendance). Taking a data-centric approach invariably leads one to ask the question how risky is the data (as a basis for classifying the data as pseudonymized personal data or anonymous information)? A more pertinent question would be – given the data how risky is the data environment? (as a basis for classify data). With that in mind, consider the dataset in respect to the following two environments – the open publication environment and the safe haven. In the case of the open publication environment such as the Internet where there are little or no infrastructure and governance controls and no restrictions on access to other data or on who can access it (which is potentially anyone in the world); the reidentification risk is likely to be very high. Given this the data should be considered as pseudonymized meaning indirectly identifying *personal data*. In the case of the safe haven environment, lets suppose that there are strict controls on both the who and how of access such as a project approval process, researcher accreditation training, physical and IT security infrastructure, and governance protocols on what can be brought into and out of the physical safe haven infrastructure; the reidentification risk is likely to be remote. Given the data and features of the safe haven environment – where the risk is determined to be remote and the agent looking at the data does not have the reasonable means to access the directly identifying information (sometimes referred to as keys) – one might reasonably make a case for classifying the data as anonymised or what Elliot et al. (2016) refer to as functionally anonymised (the concept of functional anonymization is outlined in section "Functional Anonymisation").

The Absolute Versus Risk-Based Approach

There is an ongoing debate on the role of anonymization in data protection. This debate is dominated by two approaches: the absolute approach and risk-based approach. The absolute approach has its origins in the fields of computer science and law, while the risk-based approach has its origins in the field of statistical disclosure control (SDC). In a nutshell, scholars in computer science and law have questioned the validity of anonymization – suggesting that it has failed in its purpose

(and promise to data subjects) to provide an absolute guarantee of data confidentiality. Conversely, scholars in the field of SDC have long acknowledged that anonymization is not a risk-free endeavor – and as a consequence, they have built an enormous body of work on risk assessment and disclosure control methodology for limiting reidentification risk (see, e.g., Willenborg and De Waal 2001; Duncan et al. 2011; Hundepool et al. 2012). Both approaches, it is fair to say, have influenced academic and legal literature and indeed practice in data protection. These approaches are considered albeit briefly next.

From the 1990s onward, there were a series of, what are referred to as, "reidentification demonstration attacks" by computer scientists, the purpose of which had been to illustrate that reidentification was possible in datasets considered to be anonymous (e.g., Sweeney (1997), the AOL (Arrington 2006), Netflix (CNN Money 2010) and New York Taxi case (Atokar 2014), also surname inference and identification of genomic data by Gymrek et al. 2013). One of the most famous demonstration attacks was carried out by Latanya Sweeney in 1996; she identified a Massachusetts Governor, Governor Weld, in a confidential public release hospital insurance dataset. Sweeney matched the hospital insurance dataset with a voter registration file, which she had brought for $20, matching on date of birth, gender, and zip code. She had in addition "other external information" from media reports as the Governor had recently collapsed at a rally and been hospitalized (see Barth-Jones 2016). This case, as with other more recent examples, has led to the questioning of the validity of anonymization to do what it had been expected to do, i.e., to ensure that confidential data remain confidential (Rubinstein 2016). Paul Ohm a US law professor discusses at length the AOL, Netflix, and Governor Weld reidentification demonstrations, suggesting in his 2010 paper, titled *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, that reidentification could occur with relative ease. Ohm argued that data could be useful or anonymous, but not both; he said, "no useful database can ever be perfectly anonymous, and as the utility of the data increases, the privacy deceases" (Ohm 2010: 1706). Ohm's position on anonymization has been widely critiqued (for a comprehensive critique see Elliot et al. 2018; Barth-Jones 2012, 2015, 2016). One of the key points made by those critiquing Ohm's position was that in the reidentification cases such as Governor Weld, AOL, Netflix, and New York Taxis, the data were poorly protected compared to modern standards of anonymization. More especially, the approach used in some of the cases was more akin to pseudonymization (as described in this chapter) rather than that of anonymization. It is important to recognize that just because one claims that data is anonymous does not mean that it is.

What Should We Expect: Risk and Utility

The notion of absolute or irreversible anonymization is about an expectation of zero risk of reidentification in a confidential dataset. The argument put forward by Ohm (2010), that "no useful data can ever be perfectly anonymous," Elliot et al. (2016, 2018) note is perfectly true – but, and this is the critical issue, Ohm's argument

misses the fundamental point that anonymization is not just about data protection. It is a process inseparable from its purpose to enable the sharing and dissemination of useful data. There is after all little point in sharing data that does not represent what it is meant to represent. Low utility is problematic on two accounts: (i) if the data are of little, or no, use to users, the data controller will have wasted their time and resources on them, and there may still be a re-identification (disclosure) risk present but no justifiable use case and (ii) the data could lead to misleading conclusions, which might have significant consequences if, for example, the data are used for making policy decisions (Elliot et al. 2016).

Anonymization is essentially a risk management process. Elliot et al. suggest:

> *'anonymised' (should be) understood in the spirit of the term 'reinforced' within 'reinforced concrete. We do not expect reinforced concrete to be indestructible, but we do expect that a structure made out of the stuff will have a negligible risk of collapsing'.* (Elliot et al. 2016: 1)

The role of the data controller, when anonymization is understood in this way, is to produce useful data for the intended audience(s) while managing risk such that it is remote. It is not a matter of utility versus data privacy as Ohm (2010) had suggested (i.e., as . . .“utility of the data increases, the privacy deceases,” Ohm 2010: 1706); you can have both. Achieving low risk and high utility however requires considering data in relation to its environment when assessing and managing reidentification risk.

A risk-based approach to anonymization has broad agreement, among academics at least (Rubinstein 2016), and is implemented in practice by National Statistical Institutes around the world. For example, the UK's Office for National Statistics carries out an enormous amount of work in the area of statistical disclosure risk assessment and control, to provide confidential census data in varying formats to a wide range of audiences. It is also a position supported by the UK's Statutory Authority, the Information Commissioner's Office, which stated in its 2012 Code of Practice on Anonymisation:

> *The DPA (1998) does not require anonymisation to be completely risk free – you must be able to mitigate the risk of identification until it is remote. If the risk of identification is reasonably likely the information should be regarded as personal data - these tests have been confirmed in binding case law from the High Court.* (2012: 6)

Although not explicitly stated in the legislation, it would seem that a risk-based approach is supported by GDPR. Recital 26 stipulates:

. . . To determine whether a natural person is identifiable, account should be taken of ***all the means reasonably likely to be used****, such as singling out, either by the controller or by another person* ***to identify*** *the natural person directly or indirectly. To* ***ascertain whether means are reasonably likely*** *to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments*. (GDPR 2016/679)

Note the conditioning for determining identifiability is on the *means reasonably likely to be used* to identify, not on the risk of identification. It is not within the realms of this chapter to discuss this point further, other than to say it places a spotlight on the issue of how (the *means* by which) identification might occur.

Assessing and Managing Reidentification Risk

In addition to the different perspectives on the role of anonymization in data protection, there are also differing perspectives, as introduced in section "What is Anonymisation," on the way in which the reidentification problem is understood and addressed: they are the data-centric approach and environment-centric approach. The data-centric approach is the dominant position; it sees risk as originating from and contained within data. This approach undoubtedly underpinned, for example, the Governor Weld's case, whereby those that released the hospital insurance database had failed to take account of the environment in which they were releasing the data, namely, an open environment with no governance or infrastructure controls, where potentially a large number of people could access the data and where other data existed that could be linked to it (i.e., the voter register and media coverage about the Governor's health as a high-profile figure). The focus on data comes at the expense of other wider and key considerations such as how, or why, a reidentification might happen or what skills, knowledge, or other data a person would require to ensure his or her attempt was a success (Mackey and Elliot 2013). In contrast, in the environment-centric approach, the *data environment perspective* seeks to address the limitations arising from a preoccupation with the data in question.

Data Environment Perspective

The data environment perspective has been developed from work undertaken over a number of years (Mackey 2009; Elliot et al. 2010, 2011a, b; Mackey and Elliot 2011, 2013; Elliot and Mackey 2014). This perspective posits that you must look at both the data and environment to ascertain realistic measures of risk and to develop effective and appropriate risk management strategies. Its component features are described thus:

- **Other data** is any information that could be linked to the data in question, thereby enabling reidentification. There are four key types of other data: personal knowledge, publicly available sources, restricted access data sources, and other similar data releases.
- **Agents** are those people and entities capable of acting on the data and interacting with it along any point in a data flow.
- **Governance processes** denote how agents' relationships with the data are managed. This includes formal governance, e.g., data access controls, licensing arrangements, and policies which prescribe and proscribe agents' interactions

and behavior through norms and practices, for example, risk aversion, culture of prioritizing data privacy or not, etc.
- **Infrastructure** denote how infrastructure and wider social and economic structures shape the data environment. At its narrowest level, infrastructure can be best thought of as the set of interconnecting structures (physical, technical) and processes (organizational, managerial) that frame and shape the data environment. At its broadest level, infrastructure can be best thought of as those intangible structures, such as political, economic, and social structures, that influence the evolution of technologies for data exploitation, as well as data access, sharing, and protection practices.

The data environment perspective leads to a particular anonymization approach, that of functional anonymization.

Functional Anonymization

Functional anonymization was a term first coined by Dibben et al. (2015) and later applied and developed by Elliot et al. (2016) in the *Anonymisation Decision-Making Framework* book and further discussed in the Elliot et al. (2018) paper, *Functional Anonymisation: personal data and data environment*.

Functional anonymization (FA) is, essentially, a form of anonymization which posits that one must consider both data and environment to determine realistic measures of risk and to manage that risk through reconfigurating data and or the environment. In the case of an open public release environment where the environment is pre-determined, the approach directs one to identify the need to either rethink the dissemination environment (is it appropriate) or rethink what you do to the data (i.e., further mask or alter it), to manage risk. The Anonymisation Decision-making Framework is a decision-making tool for doing functional anonymization, which guides one to consider how the law, ethics, and social expectations may interact with the process of anonymization.

Anonymisation Decision-Making Framework

The Anonymisation Decision-making Framework was developed by Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. It represents a broader collaborative piece of work undertaken by the UK Anonymisation Network (UKAN) of which the authors of the ADF are founding members. The ADF's broader collaborative underpinnings come about from the input of UKAN's core network of 30 representatives (drawn from the public, private and charity sectors) in considering 2 fundamental anonymization questions:

1. How should anonymization be defined and described, given the many different perspectives on it?

2. What should practical advice look like, given that anonymization is a complex topic requiring skill and judgment?

The core network's input, in to addressing these questions, was captured in a series of workshops. The ADF developed, thereafter, was a ten-component framework, spanning three core anonymization activities, that of, (i) data situation audit, (ii) risk assessment and management, and (iii) impact management. The purpose of developing the ADF was to fill the gap between guidance given in the ICO's 2012 Code of Practice on Anonymisation, and that which is needed when grappling with the practical reality of doing anonymization.

In early 2018, with the advent of the General Data Protection Regulation UKAN undertook a further engagement exercise (over a 6-month period) meeting with legal and privacy experts from the UK and Europe, and UKAN's user community, to consider two issues:

1. The likely impact of GDPR on the ADF
2. How the framework had so far been received and applied in practice, since its publication in 2016

As a result of this engagement work, a variety of new materials are being developed including a second edition of the ADF book. In providing an overview of the ADF, in this chapter material from the 2016 publication and from more recent developments has been drawn on. It is worth noting that work on the second edition of the ADF is ongoing and not yet published; the details of the component framework given here may differ to that which is given in the second edition – however the essence of what is being written about will not.

Anonymisation Decision-Making Framework

The framework is founded on four principles:

- **Comprehensiveness principle**: posits that one cannot determine identification risk by examining the data alone (the data-centric approach) and you must consider both data and environment (data environment perspective).
- **Utility principle**: posits that anonymization is a process to produce safe data, but it only makes sense if what you are producing is safe useful data.
- **Realistic risk principle**: is closely associated with the utility principle; it posits that zero risk is not possible if one is to produce useful data. Anonymization is about risk management.
- **Proportionality principle**: posits that the measures you put in place to manage reidentification risk should be proportional to the likelihood and (likely) impact of that risk.

The ADF has, since it was first written about, evolved into a more nuanced 12-component framework, and those components are:

1. Describe the use case.
2. Sketch the data flow.
3. Map the properties of the environment(s).
4. Describe the data.
5. Map the legal issues.
6. Meet your ethical obligations.
7. Evaluate the data situation.
8. Select the processes you will use to assess and control risk.
9. Implement your chosen approach to controlling risk.
10. Continue to monitor the data situation.
11. Plan how to maintain trust.
12. Plan what to do if things go wrong.

These components cover the same three anonymization activities noted previously: data situation audit, risk assessment and risk management, and impact management. The framework can be used to: (i) establish a clear picture of one's processing activities and (ii) assess and manage risk. Anonymization however is not an exact science; one will still need to make judgment calls as to whether data are sufficiently "anonymized" for a given data situation (Elliot et al. 2016).

Data Situation Audit

The term **data situation** is used to capture the notion of data in an environment. A data situation can either be static or dynamic. Static data situations are where data exist within a single (closed) environment – no data in or out. Most situations are however dynamic situations where data moves because it is shared internally and/or externally.

A data situation audit can be undertaken as:

- A standalone piece of work, to provide an audit of one's processing activities and to demonstrate compliance with the GDPR (2016) and DPA (2018).
- It can be used to feed into the anonymization activity of risk assessment and control.

In presenting the components of the ADF in this chapter for ease of illustration, a simple share scenario is used involving the flow of data between three environments; data flows are commonly more complex than this.

Component 1: Describe the Use Case

The use case is principally the rationale for a data share (internally or externally) or release. It is likely to be a strong determent of what data is shared, or released, to whom, and by what means, so it is important to establish what the use case is early on. The use case is intrinsically connected to decisions about the utility-risk balance.

The aim, of course, is to provide high-utility low-risk data that meet the requirement of the intended audience.

Component 1 is about considering the why, who, and how associated with a use case, by identifying:

1. The rationale for sharing or releasing data.
2. Who the groups are, that may want to access the data being shared or released.
3. How those accessing the data might want to use it.

Hypothetical Scenario of a Data Share

Let us imagine that a research team (based at University A) applies for an extract of motor vehicle accident data held by the Department of Health & Safety (DHS). The DHS agrees to share the data extract; under the terms of a data sharing agreement the DHS will securely transfer the data extract to a safe haven at University A. In the data sharing agreement the DHS also stipulate for what purpose the research team can analyze the data and the conditions for publishing research results (i.e. research outputs must be checked against ESSNet guidelines to ensure they are not disclosive). The purpose of the research is to better understand the factors and impact associated with vehicle accidents. This scenario is used as a basis for outlining the other components.

Component 2: Sketch the Data Flow

Sketching a data flow from the origin of data collection across the environments in which it will be held allows one to visualize the parameters of a data situation.

Hypothetical Scenario of a Data Share

Figure 1 illustrates the data flow between DHS, the safe haven environment, and publication environment.

Component 3: Map the Properties of the Environment(s)

It is important to describe, in a structured way, data environments taking into account, other data, agents, infrastructure, and governance.

This information feeds into:

- Component 7: to help evaluate the data situation
- Component 8: if further work on risk assessment and management is required

Hypothetical Scenario of a Data Share

University A, acts as the Data Controller for the data extract provided to it; in order to fully understand the risks associated with processing the data the research team needs to establish what the properties of the share and publication environments are. Lets imagine that the safe haven environment is as described in Table 1, and that the research publication environment is as described in Table 2.

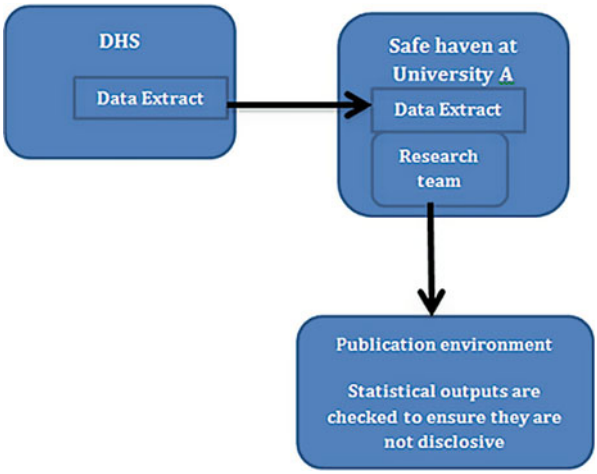

Fig. 1 Data flow between DHS, University A "safe haven," researcher, publication environment

Table 1 Safe haven environment

Data to be shared	Extract of data from DHS
Other data	No unauthorized data can be brought into or removed from the safe haven All research outputs checked to ensure they are not disclosive
Agents	Research team at University A Researchers must have completed accreditation training course
Infrastructure	Restrictions placed on who can access the safe haven Secure IT infrastructure (ISO 27001 compliant) Controls on the work-space
Governance processes	User Agreement between research team and Safe Haven SOP on how to work in the safe haven and penalties for misuse of safe haven

Table 2 Research publication environment

Data to be shared	Aggregate data derived from the extract of shared data and checked against ESSNet guidelines
Other data	Potential data sources in the pubic domain include: 1. Publicly accessible sources, e.g., public records, social media data, registers, Newpaper archives etc. 2. Personal knowledge 3. Other similar datasets 4. Restricted access datasets
Agents	Potentially anyone in the world
Infrastructure	Open, few infrastructure controls
Governance processes	Open, no governance controls

Table 3 Hypothetical data share – DHS–University A: data risk profile

Risk profile	Proposed extract of shared data
The data structure	Numerical
The data type	Microdata
The data population type	Population data
The variable types	**No direct identifiers included** **Quasi-identifiers:** accident type, date of accident (month and year), number of vehicles involved, number of persons involved, number of fatalities, number of persons with life changing injuries, road conditions, accident location, demographics of persons involved: age in single years, gender **Special category data**: ethnic origin, concomitant health, concomitant medication
The dataset property type	Cross-sectional, large-scale dataset; age of data - 2 years old; hierarchical data - yes; family relationships captured; quality of data - good
Topic area type	Sensitive

Component 4: Describe the Data

As well as understanding the data environment(s) you need to understand the data, one way of doing this is to create a risk profile for data.

A risk profile can be established by specifying:

- **The data structure:** i.e., whether the data in question is numerical, text, film, an image, etc.
- **The data type:** i.e., whether the data is microdata (at the individual level) or aggregate data (at the group level).
- **The data population type:** i.e., whether the data represents a sample or whole population. A whole population may be a census or all people in a particular population, such as all passengers on a shipping manifest.
- **The variable types:** i.e., whether there are any special features, such as direct identifiers, quasi-identifiers, and special category data.
- **The dataset property type:** this includes identifying relationships in the data structure (e.g., households), data quality, size of dataset, age of data, and time period captured (e.g., one off or overtime in the case of longitudinal data).
- **Topic area type:** i.e., whether the topic area may be considered sensitive.

Hypothetical Scenario of a Data Share

Let us imagine that the extract of shared data provided by DHS to the research team at University A has the following *risk profile* (given in Table 3).

A risk profile is a top level assessment of the data which feeds into:

- Component 7: to evaluate the data situation
- Component 8: if further work on risk assessment and management is required

Table 4 Hypothetical data share: DHS–University A legal issues profile

Hypothetical data share	DHS-University A
Provenance of the data What one needs to know is where the data originates because this will help in determining processing responsibilities	DHS collates accident data across the UK from data provided by local Police Forces. DHS is the data controller for this dataset.
Means and purpose of processing	University A is considered a data controller for the extract of shared data because the research team has determined the means and purpose of the processing in this scenario. University A's responsibilities as a joint data controller for the extract of shared data are agreed in a contract between DHS and University A.
GDPR considerations	The research team's legal basis for processing the accident data is: **Article 6** (1e) – Public task **In addition for processing special category data, Article 9** (2j) – for scientific and research purposes As the proposed share is new and is on a sensitive topic area, a Data Protection Impact Assessment will he carried out by the research team; this will provide a clear audit trail of University's A processing activities and ensure that a data protection by design approach is embeded in the project at the planning stage.
Mechanism for sharing	Data sharing agreement between DHS and University A

Component 5: Map the Legal Issues

The movement of data across multiple environments can complicate the issue of roles and responsibilities. Establishing a profile of the legal issues can help clarify who is the data controller, data processor and data user. The profile can be established by specifying:

- **The provenance** of the data
- **Who has determined the purpose** for which, and the manner in which, personal data is processed
- **GDPR requirements**, i.e., a legal basis for processing, whether a data protection impact assessment is needed, etc.
- **Other relevant legislation** enabling the share (E.G. Part 5 of the Digital Economy Act, 2017; Common Law duty of confidence)
- **The method used for formalizing the share**, e.g., a data sharing agreement, contract, or licensing agreement

Hypothetical Scenario of a Data Share

Let us imagine that the data extract shared between DHS and University A has the following profile (described in Table 4*, Legal issues Profile*).

Table 5 Hypothetical data share – DHS–University: Assessment of data situation sensitivity

1.	Is the proposed purpose for which the data will be reused likely to be considered inconsistent with the original data collection purpose? **No**
2.	Is the planned data share new? **No**
3.	Does the project involve a commercial sector partner, investor, or benefactor? **No**
4.	If a commercial sector partner is involved, will they have access to the data? **No**
5.	Are special category data being accessed? **Yes**
6.	Have data subjects consented to the proposed reuse of the data? **No**
7.	Has any public engagement activities been planned? **Yes**
8.	Is information about the project clear and accessible? **Yes**

Evaluation:
The more you answer yes to questions 1–5 and no to questions 6–8, the greater the data situation sensitivity: **Given the answers 1–7, the level of data situation sensitivity is considered relatively low**

Component 6: Meet Your Ethical Obligations

The reason why ethics is an important consideration is that anonymization is a process that invariably involves the processing of personal data; even once data has gone through the anonymization process there are reasons for thinking about ethics, this is because: (i) data subjects might not want data about them being reused in general by specific third parties or for particular purposes and (ii) the risk of a confidentiality breach is not zero. The notion of data situation sensitivity is key here; underpinning this concept is the idea of reasonable expectations about data use and or reuse – more especially the need not to violate (data subjects and the publics') expectations about how data can be used. An assessment of a data situation's sensitivity can be carried out by answering the following questions:

1. Is the proposed purpose for which the data will be reused likely to be considered inconsistent with the original data collection purpose?
2. Is the planned data share new?
3. Does the project involve a commercial sector partner, investor, or benefactor?
4. If a commercial sector partner is involved will they have access to the data?
5. Are special category data being accessed?
6. Have data subjects consented to the proposed use/reuse of the data?
7. Has any public engagement activities been undertaken?
8. Is information about the proposed data processing clear and accessible to key stakeholders?

The more you answer *yes to questions 1–5* and *no to questions 6–8*, the greater the likelihood of a potentially sensitive data situation (Mackey and Thomas 2019).

Hypothetical Scenario of a Data Share

Let us imagine that the data situation sensitivity for the extract of shared data is as described in Table 5, *Assessment of Data Situation Sensitivity.*

Component 7: Evaluate the Data Situation

The parameters of a data situation can be determined by mapping the expected flow of data across the environments involved in the data share or dissemination. The data flow can be populated with the key relevant information across the data flow, i.e., the use case and sensitivity profile, and also within each environment, i.e., data risk profile, description of environment and legal issues and data sensitivity profile. This evaluation should feed into the next anonymization activity risk assessment and risk managment.

Risk Assessment and Risk Management

Component 8: Select the Processes You Will Choose to Assess and Control Risk

A formal assessment may include all three parts of the process described below or only some elements of it.

1. An analysis to establish relevant plausible scenarios for the data situation under review. Scenario analysis considers the how, who, and why of a potential breach.
2. Data analytical approaches to estimate risk, given the scenarios developed under procedure 2.
3. Penetration testing to validate assumptions made in procedure 2, by simulating attacks using "friendly" intruders.

The first procedure is always necessary, while the second and third may or may not be required depending on the conclusions drawn from the previous two.

Component 9: Implement Your Chosen Approach to Control Risk

Processes for controlling risk essentially attend to either or both elements of a data situation: the data and their environment. If the risk analysis in component 8 suggests that stronger controls are necessary, then there are two non-exclusive options:

1. Change the data specifications.
2. Reconfigure the data environment.

Component 10: Continue to Monitor the Data Situation

Once you have shared or disseminated data you need to continue to monitor the risk associated with those data over time. This may involve monitoring technological developments and the availability of other data, as well as keeping a register of all data shared to cross-reference with future releases.

Impact Management

Much of what has been set out thus far has been framed in terms of risk management, but one should in addition prepare for if the worst should happen. Impact management is all about putting in place a plan for reducing the impact of a reidentification should it happen.

Component 11: Plan How to Maintain Trust

Effective communication with key stakeholders is crucial to build trust and credibility, both of which are critical to difficult situations, where a data controller might need to be heard, understood, and believed. The key point is that a data controller is better placed to manage the impact of a reidentification, if they and their stakeholders have developed a good working relationship.

Component 12: Plan What to Do if Things Go Wrong

If, in the rare event, a breach were to occur. It is recommended that:

1. All processing activities are clearly document to provide a clear audit trail. This is in fact a requirement under GDPR (2016/69) which has newly introduced the principle of accountability.
2. Plans and policies are developed and communicated to ensure it is clear who should do what, when, and how if a breach occurs.

Concluding Remarks

This chapter provides an introduction to the topic of anonymization: first outlining the key terminology of pseudonymization and anonymization. Pseudonymization is described as addressing no more than the risk arising directly from data, while anonymization should address the risk arising both directly and indirectly from data. To classify and address identification risk arising directly and indirectly from data, it has been argued, in this chapter, that one must take account of data in relation to the environment in which it exists. Thus anonymization is defined as, *a process whereby personal data, or personal data and its environment, are modified in such a way as to render data subjects no longer identifiable.*

The discussion then moves on to illustrate how one thinks about the reidentification problem influences how one applies the concepts of pseudonymization and anonymization. There are two approaches to thinking about the reidentification problem, the data-centric and environment-centric approaches. The data-centric approach is the traditional and dominant approach – it focuses on data at the expense of key considerations associated with the who, how, and why of reidentification. In contrast the data environment approach considers data in relation to its environment(s) to take account of the who, why, and how of reidentification. It is this latter approach that informs and underpins the Anonymisation Decision-making Framework.

Just as there are two perspectives on how to understand and address the reidentification problem, there are two approaches on the role of anonymization in data protection. These two approaches are the absolute approach that makes the argument that anonymization should be irreversible and the risk-based approach that makes the argument that there is an inherent risk of reidentification in all useful data and so the role of the data controller is to ensure that the risk is remote. It is the latter approach that underpins the Anonymisation Decision-making Framework.

Anonymization should be understood as a process inseparable from its purpose of analysis and dissemination which means it only makes sense if what is produced is *useful* anonymised data. It is possible to have both high utility and low risk data if one considers data in relation to its environment(s). Anonymization understood in this way, Elliot et al. (2016, 2018) call functional anonymization. The ADF is a tool for doing functional anonymization.

In the final section of the chapter, a decision-making tool for doing well-thought-out anonymization was introduced, the Anonymisation Decision-making Framework. The ADF unifies the legal, ethical, social, and technical aspects of anonymization into a single framework to provide best practice guidance.

Cross-References

- ▶ Big Data
- ▶ Biosecurity Risk Management in Research
- ▶ Creative Methods
- ▶ Data Sharing and Data Archiving
- ▶ Privacy
- ▶ Research Ethics in Data, Data Protection, Transfer and Urbanism
- ▶ Research Governance

References

Arrington M (2006) AOL proudly releases massive amounts of user search data. TechCrunch. http://tinyurl.com/AOL-SEARCH-BREACH. Accessed 30 May 2016

Atokar (2014) Riding with the stars: passenger privacy in the NYC taxicab dataset. http://tinyurl.com/NYC-TAXI-BREACH. Accessed 30 May 2016

Barth-Jones D (2012) The identification of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. https://fpf.org/wp-content/uploads/The-Re-identification-of-Governor-Welds-Medical-Information-Daniel-Barth-Jones.pdf

Barth-Jones D (2015) How anonymous is anonymity? Open data releases and re-identification. Data & Society. https://datasociety.net/pubs/db/Barth-Jones_slides_043015.pdf

Barth-Jones D (2016) why a systems-science perspective is needed to better inform data privacy public policy, regulation and law. Brussels privacy symposium, November 2016

CNN Money (2010) 5 data breaches: from embarrassing to deadly. http://tinyurl.com/CNN-BREACHES/. Accessed 30 May 2016]

Concordat on Open Research Data (2016). https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/

Dibben C, Elliot M, Gowans, H, Lightfoot D, Data Linkage Centres (2015) The data linkage environment. In: Harron K, Goldstein H, Dibben K (ed) Methodological Developments in Data Linkage, First Edition. Edited by Katie Harron, Harvey Goldstein and Chris Dibben. © 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd

Duncan GT, Elliot MJ, Salazae-Gonzalez JJ (2011) Statistical confidentiality. Springer, New York

Elliot M, Mackey E (2014) The social data environment. In: O'Hara K, David SL, de Roure D, Nguyen CM-H (eds) Digital enlightenment yearbook. IOS Press, Amsterdam

Elliot M, Lomax S, Mackey E, Purdam K (2010) Data environment analysis and the key variable mapping system. In: Domingo-Ferrer J, Magkos E (eds) Privacy in statistical databases. Springer, Berlin

Elliot M, Smith D, Mackey E, Purdam K (2011a) Key variable mapping system II. In: Proceedings of UNECE worksession on statistical confidentiality, Tarragona, Oct 2011

Elliot MJ, Mackey E, Purdam K (2011b) Formalizing the selection of key variables in disclosure risk assessment. In: 58th congress of the International Statistical Institute, Aug 2011, Dublin

Elliot M, Mackey E, O'Hara K, Tudor C (2016) The anonymisation decision-making framework. UKAN Publication, Manchester, United Kingdom

Elliot M, O'Hara K, Raab C, O'Keefe C, Mackey E, Dibben C, Gowans H, Purdam K, McCullagh K (2018) Functional anonymisation: personal data and the data environment. Comput Law Secur Rev 34(2):204–221

ESSNet (2007) Guidelines for the checking of output based on microdata research, Workpackage 11. Data without Borders. Project N°: 262608. Authors: Steve Bond (ONS), Maurice Brandt (Destatis), Peter-Paul de Wolf (CBS). Online at https://ec.europa.eu/eurostat/cros/content/guidelines-output-checking_en

Fienburg SE, Makov UE, Sanil A (1997) A Bayesian approach to data disclosure: optimal intruder behaviour for continuous data. J Off Stat 13(1):75–89

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. Science 339(6117):321–324. https://doi.org/10.1126/science.1229566. [PubMed]

Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, DE Wolf PP (2012) Statistical disclosure control. Wiley, London

ICO Anonymisation: managing data protection risk code of practice 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf

Mackey E (2009) A framework for understanding statistical disclosure control processes. PhD thesis, The University of Manchester, Manchester

Mackey E, Elliot M (2011) End game: can game theory help us explain how a statistical disclosure might occur and play out? CCSR working paper 2011–02

Mackey E, Elliot M (2013) Understanding the data environment. XRDS 20(1):37–39

Mackey E, Thomas I (2019) Data protection impact assessment: guidance on identification, assessment and mitigation of high risk for linked administrative data. Report for the Administrative Data Research Partnership

Mourby M, Mackey E, Elliot M, Gowans H, Wallace S, Bell J, Smith H, Aidinlis S, Kaye J (2018) Anonymous, pseudonymous or both? Implications of the GDPR for administrative data. Comput Law Secur Rev 34(2):222–233

Ohm P (2010) Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 57(1701):1717–1723

Open Research Data Taskforce with Michael Jubb (2017) Research data infrastructure in the UK landscape report. https://www.universitiesuk.ac.uk/policy-and-analysis/research-policy/open-science/Pages/open-research-data-task-force.aspx

Open Research Data Taskforce (2018) Realising the potential. Open Research Data Taskforce final report. https://www.gov.uk/government/publications/open-research-data-task-force-final-report

Rubinstein I (2016) Brussels Privacy Symposium on Identifiability: policy and practical solutions for anonymisation and pseudonymisation – framing the discussion. In: Proceedings of Brussels Privacy Symposium: identifiability: policy and practical solutions for anonymisation and pseudonymisation. Brussels, Nov 2016. https://fpf.org/wp-content/uploads/2016/11/Mackey-Elliot-and-OHara-Anonymisation-Decision-making-Framework-v1-Oct-2016.pdf

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free

movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Online at https://eur-lex.europa.eu/legalcontent/EN/TXT/?qid=1568043180510&uri=CELEX:32016R0679

Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. J Law Med Ethics 25(2–3):98–110. https://doi.org/10.1111/j.1748-720X.1997.tb01885.x

UK Data Protection Act (2018) London, The Stationery Office. Online at http://www.legislation.gov.uk/ukpga/2018/12/contents/data.pdf

Willenborg L, DE Waal T (2001) Elements of disclosure control. Springer, New York