Functional Anonymisation and the Role of the Data Environment in Determining the Classification of Data as (Non-)Personal

### Abstract

This paper focuses on the extent to which the environment within which data exist (their *data environment*) can make data – which would otherwise be deemed personal – non-personal; this distinction being central to much data protection and privacy law globally. Such an environment would necessarily ensure that the elements required to re-identify the data will not be present or, conversely, that the features of the environment are sufficient to prevent such re-identification. The paper proposes that: (i) the personal or non-personal status of data is dependent on that data’s environment; (ii) the status of being non-personal is isomorphic with it being *functionally anonymised*;(iii) the data environment can be defined by a small number of parameters including: other data in the environment, the skills and knowledge of the persons who are present in the environment, governance structures and process, and infrastructure; (iv) those parameters are controllable; and (v) it is possible for a significant proportion of useful, previously considered personal datasets to achieve a status of *environment dependent non-personal data.* We provide a formulation for describing the relationship between the data and its environment which links together the legal notion of personal data with the statistical notion of disclosure control. An assessment of the data environment and the functional anonymisation approach gives data providers, custodians and users an optimum framework for accessing and using data and so can make a valuable contribution to the knowledge economy.

# Introduction

No data exists in isolation. In this paper we refer to the context in which data exists as its *data environment* (Mackey and Elliot 2013). The paper proposes that the properties of a data environment can make the data within it – which would otherwise be deemed personal – non-personal. Releasing data that can be used to identify an individual is not only a risk to the individual concerned but also the data release programme and the data holders. In order to facilitate safe and legal data sharing, it is necessary to release data into an environment in which the elements required to re-identify the data and/or the motivations to do so will not be present, or alternatively, to construct such an environment.

Here we propose that in order to determine whether or not particular data are personal or non-personal for the purposes of data protection and indeed other legislation, the data environment must first be assessed. Indeed, the notion of the *data environment* is now accepted as part of the interpretation of the *Data Protection Act (1998)* and other instruments in the UK.[[1]](#endnote-1) However, as both Bonnici (2014) and Wright and Raabb (2014) observe, the practice of incorporating such assessment into decisions about whether data are personal or not is poorly defined and problematic.

This paper aims to address these issues and proposes that:

1. The personal or non-personal status of data is dependent on that data’s environment.
2. The data environment can be understood by a small number of parameters including: other data in the environment, the skills, knowledge and motivations of the persons who are present in the environment, governance structures and process, and infrastructure.
3. Those parameters are to a large extent controllable.
4. It is possible for a significant proportion of useful datasets to achieve a status of *environment dependent non-personal data.*

To develop this argument we will first define the key terms “Personal Data”, “Anonymisation” and “Data Environment”. We will then show how parameters of the data environment can alter the interpretation of the data as personal or not.

# Personal or Non-Personal Data

The concept of personal data derives from the principles of *data protection* which are enacted in legislation in jurisdictions across the world. The European Union (EU) data protection directive 95/46/EC – soon to be superseded bythe General Data Protection Regulation (GDPR) – is enacted through national legislation in the EU’s member states. The legal obligations and duties that such data protection law impose upon the holders of data (known as *data controllers*) are effectively constraints on ‘the processing of personal data’ even though the right to personal data is not absolute (for discussion see Bonnici 2014).

Yet the directive is not quite clear as to how to classify data as personal. Article 2(a) of the directive defines personal data as “any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.”

This characterisation immediately raises the obvious question: ‘identifiable by whom?’ After all, there will always be someone who could identify a data subject (quite often, the data subject him- or herself). Moreover, depending on whether the answer to that question is taken broadly or narrowly, so the scope of data protection will be broad or narrow.

In an EU directive, the articles are enacted in national legislation. However, these are preceded by a commentary, expressed in recitals, which advise governments but which do not have to be enacted. Recital 26 of 95/46/EC states among other things that “the principles of protection must apply to any information concerning an identified or identifiable person” and that “to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.” This is a very broad definition, which answers the ‘by whom’ question as widely as possible – by anyone.

However, in the introductory section of the UK’s *Data Protection Act (1998)*, personal data is defined as “information or data which relate to a living individual who can be identified from those data, or from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller (including any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect to the individual).” In other words, the UK Act is consistent with the articles of the directive, but falls short of the specification set out in the recitals at the beginning. It deliberately narrows the scope of the legislation by answering the ‘by whom’ question with the single instance of the data controller.

This is not the only approach to privacy protection. For instance, in the United States, what is called personally identifying information (PII) is defined piecemeal in three types of legal context. One is privacy law, where PII is defined abstractly, as in the EU directive. This produces a patchwork of partial definitions. The *Privacy Act* of 1974 regulates the federal government’s collection of personal data, while several states also have their own generic privacy laws. No federal law governs the use of data by private-sector bodies. The second context is that of sector-specific privacy law, such as the Health Insurance Portability and Accountability Act (HIPAA), which defines PII as information which “1) identifies the individual; or 2) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.” However, HIPAA applies only to health information. Finally, there are various breach notification laws, which order organisations which have leaked PII to notify the data subjects. These tend to define specific data types, so for example California’s breach notification law defines PII as including Social Security numbers, driver’s license numbers and financial accounts, but not email addresses or telephone numbers.

For some data – often called formal identifiers – it is fairly straightforward to determine that they are personal. For example, an individual’s name, address, date of birth, bank account numbers or debit or credit card details all constitute clear-cut ‘personal data’. However, for other types of data determining whether it is ‘personal’ or ‘non-personal’ can be more complicated. For instance, if a group of data subjects is known to contain only one woman, then the gender of data subjects (not normally considered an identifying feature) will be identifying for the woman. Identifiers can also be constructed out of combinations of attributes (for example, consider a “sixteen year old widow” or a “15 year old male University Student” or a “female Bangladeshi bank manager living in Thurso” (see Elliot and Dale 1999 for discussion of this issue).

The decision about whether data is personal or non-personal is particularly important for research and business use. In order to process any personal data fairly and lawfully, the *Data Protection Act (1998)* requires a Schedule 2 condition to be satisfied, and a further Schedule 3 condition to be satisfied in relation to sensitive personal data. Other statutory legal provisions placed upon specific data controllers may also have conditions of use of personal data and must be understood in parallel with the *Data Protection Act*. This minefield of legal obstacles can be difficult to understand and is both inhibitive and restrictive towards the use of personal data for meaningful research.

If, however, data is deemed non-personal then data protection legislation does not apply.[[2]](#endnote-2) This is potentially of benefit to the data controller who will then be able process the data for purposes other than originally collected and maximise its research and economic value.[[3]](#endnote-3) The innovation of this paper, that the environment in which data is held affects its personal or non-personal status, is therefore crucial for data use.

For the purposes of data protection legislation, non-personal data is the converse of personal data, i.e. it is information which does not contain any information relevant to identifying living individuals and, more importantly for our current purposes, it could be personal data that has been successfully anonymised.

It is noteworthy that whatever approach is developed to define personal data, it invariably revolves around properties of the data. In particular, if data of a particular group of types (counted as personally identifying) is present in the dataset, then that is regarded as sufficient to mark it as personal data. Context is abstracted away, sometimes by fixing it unrealistically. Indeed, any fixed answer to the ‘by whom’ question will unrealistically determine a single context. However, it is the contention of the current paper that this approach, abstracted as it is away from context, will prove insensitive to the weight of risk in sharing a dataset. When assessing privacy risks, it makes much more sense to consider the ways in which data are likely to be used, and the context into which it will be released. Properties of the data are of course important, but even apparently innocuous data can be dangerous when released into a very rich context. When we talk about the ‘data environment’ below, our aim is to provide a language for describing this context, and its relevant aspects.

# Understanding Data Anonymisation

In terms of its practical application the key issue in the definition of personal data is the notion of identifiability. If a data subject cannot be identified the data is non-personal. But how do we determine whether a data subject can be identified or not? The notion of identifiability is related to another concept: *anonymisation*. The mapping here is if a dataset is anonymised then it is also non-personal. To confuse matters somewhat, the term “Anonymisation” is polysemous and used in a least three ways: Formal, Absolute and Statistical. We argue here that the first two are not useful in any practical sense and the last focuses too heavily on the statistical properties of the data. A fourth concept which we introduce here is *functional anonymisation*.

#### Formal Anonymisation

For data to be formally anonymised simply requires that *formal identifiers* have been removed from the data set. Formal identifiers come in three forms:

1. Intentional Unique Identifiers. These are serial numbers that have been created with the explicit intention of identifying a person and for linking transactions. They are often used in multiple contexts and usually are associated with a person across his or her lifespan. In the context examples are: UK National Insurance Numbers and US Social Security Numbers.
2. Transactional Unique Identifiers. These are numbers which have been generated as part of some transactional process. They are not necessarily permanent. The most obvious example is a telephone number – particularly a mobile phone number.
3. Functional Unique Identifiers. This category is a borderline one and it is certainly debatable whether FUI’s should count as Statistical identifiers. FUIs will almost always be constructed out of more than one piece of information. They will also usually include the possibility of rare data twins. The most straightforward of the FUI’s is full name and address. A more debateable case is the combination of postcode and date of birth.

#### Absolute Anonymisation

For data to be absolutely anonymised there must be zero risk of an individual being identified within it. This is the meaning of anonymisation that is usually employed within the security engineering (see Ohm 2010; Dwork 2006, etc.) and in particular through the application of differential privacy. The aim of differential privacy is to provide a *privacy guarantee*, through algorithms that make very specific and extreme assumptions about what a data user might already know about the population represented in the data.

The problem with Absolute Anonymisation is that in order to achieve it, one has to so restrict the data that it is often rendered useless. As Muralidhar and Sarathy (2011) demonstrate when differential privacy techniques are applied to an analysis server, the net effect is that meaningful queries to the differential private database are no longer possible. The result is not that surprising when one considers that plausible re-identification attacks and meaningful data analysis are both predicated on the same thing: data that differentiates population units.

#### Statistical Anonymisation

The notion of statistical anonymisation is tied into a technical field called *statistical disclosure control* (SDC, also sometimes called *statistical confidentiality*). The basic tenant of SDC is that it is impossible to reduce the probability of re-identification to zero. Instead of which one needs to control or limit the risk of disclosure events.[[4]](#endnote-4) This brings the notion of anonymisation into line with other areas of business risk management. One accepts that our actions and choices, responsibilities and constraints are embedded in a complex world which is impossible to predict so one gathers the very best information one can and optimises one’s decisions to maximise the benefits and minimise the risks.

It should be clear from the above that the only practical notion of anonymisation is the statistical one. Indeed one could argue that the other two notions are simply special cases. Formal anonymisation is simply a mechanism for reducing the risk of re-identification below unity and absolute anonymisation is a mechanism for reducing the probability of re-identification to zero. Given that in releasing or sharing data that relates to individuals one has two goals: to release/share useful data and to release/share data in a form which protects confidentiality (and thereby privacy), it should be clear that the other two definitions of anonymisation fail to meet one or other of these goals.

Unfortunately, assessing disclosure risk, even with the simplest of data is a far from trivial problem. So much so that a whole research community has built up around the topic with its own journals and conferences (see Duncan et al. 2011 or Hundepool et al. 2012 for recent reviews of the field). Much of the work in this field has focused on the statistical properties of the data to be released/shared. Primarily because of this aspect of the disclosure risk problem to be considered it is by far the most tractable. A great deal of headway has been made; sophisticated statistical models have been developed which at least allow identification probability assessments to be anchored in the data properties.

However, as several authors (see e.g. Paass 1988; Elliot and Dale 1999; Mackey and Elliot 2009; 2013) have pointed out we are at best here basing our measurement on only part of the components of the risk. There are a whole range of other issues:

1. The motivation of somebody wishing to attack anonymised data in order to re-identify somebody within it (this will effect *what* happens and *how*).
2. What the consequences of a disclosure are (which will affect the motivations of an individual to attempt a re-identification).
3. How a disclosure might happen without malicious intent (the issue of *spontaneous identification*).
4. How the governance structure, data security and other infrastructural properties surrounding the release/sharing of the data affect the risk.
5. The other data/knowledge that might be linked to the data in question (without which disclosure/identification is impossible).
6. Differences between the data in question and the other data/knowledge (often referred to as *data divergence*).

Bringing these considerations into the framework of statistical anonymisation creates the fourth type: *functional anonymisation*. These comprise the contextual factors which we refer to collectively as the *data environment*.

# Defining the Data Environment

The term “data environment” was first coined by Elliot et al. (2011) who originally defined it instrumentally as “the set of all possible data that might linked to a given dataset” page? . At its broadest though the data environment might be defined as the context in which any item of data exists. Here in line with the more recent definition used by Mackey and Elliot (2013) we define it thus: *the set of (formal or informal) structures, processes, mechanisms and agents that either (i) interact with given data; (ii) control interactions with that data; or (iii) provide interpretable context for that data.*

A brief examination of this definition immediately raises three interrelated questions:

1. What are the elements of a data environment – what types of things exist in it/constitute it?
2. How is the environment structured?
3. What is the relationship between the data environment and the socio-physical one?

### The elements of a data environment

Tackling the first question first we propose that a data environment usually consists of four key elements:

1. Data
2. Data users
3. Governance processes
4. Infrastructure

The first element may seem tautologous, but it is arguably the only necessary element. What makes the data environment an *environment* is that it is *teeming with data in a complex web of interrelationships*. From our own point of entry this is critical – the missing component in our decision about whether data is anonymised (and therefore personal or not) is an understanding of the information that could be linked to data that we are about to disclose (i.e. data that will be in that data’s environment once it is disclosed).

The second element, the data user, motivates and operates on/in the data environment. This is also arguably a logically necessary component, as empirically, without users there would be no reason for data and therefore no data environment.[[5]](#endnote-5) Data users capture data, move it, transform it, link it and analyse it. Through such operations the data environment is transformed but data user behaviour is also shaped by the structure and processes of the data environment. Therefore, it is reasonable to say that data users also exist in the data environment.

Another point about users is that they, as well as being agents within the data environment, also provide part of the interpretable context. This may seem a little odd but without this understanding the definition of personal data in the UK *Data Protection Act*: “....data which relate to a living individual who can be identified from those data, or from those data and other information...”[[6]](#endnote-6) becomes partially meaningless. If one interpreted the emboldened section literally it would mean identification “without any interpretable context” which is logically impossible. Clearly the implicit qualifier is something like “given the minimal interpretative context that any data user could reasonably be expected to bring to bear”.

The third element of the data environment is governance processes. We use the terms here broadly to mean how the users’ relationships with the data are managed. This includes formal governance (e.g. data access controls, licensing arrangements, and policies which prescribe and proscribe user behaviour) through de facto norms and practices to socio-cognitive user *set* (e.g. risk aversion, prior tendency towards disclosure, etc.).

The final element is infrastructure. This includes physical elements such as security systems as well as the software processes such as functional restrictions on how users can interact with the data.

### The structure of the data environment

Thus far in this section we have been talking as if there is a compartmentalised data environment in which each dataset resides. For expository and motivational purposes this was useful. However, theoretically it is unsophisticated. It should be immediately obvious that data and its local environment also exist within a global data environment: the complete data overlay of all human activity.

That global environment has a structure which comprises local data environments and the relationships between them. A key element of this structure that produces immense complexity is the various forms of vertical and horizontal *partioning*. In between the open global data environment there are a layers caused by the elements described above. National-level legislation creates governance. Organisations build IT security infrastructure to keep the rest of world out and their own data in. Within functional units there will be subdivisions, for example a company may internally restrict access to particular data. All of these constitute a mix of vertical and horizontal partitions of the global data environment (and in effect create local environments).

Complicating this picture still further are agents – the data users – who move in and out of data environments. As agents necessarily store and process information, they themselves are in effect mobile data environments that change other data environments as they enter and leave them, bringing and taking with them data-knowledge and knowledge about data.

So the global data environment is complex and fluid – and that complexity and fluidity is to some extent reflected at all levels – in all local data environments. If this were not so then the problem of making personal data non-personal could be dealt with deterministically. However, we can place operational controls on a local data environment; controls which operate on the four elements (data, users, governance, infrastructure) and these controls have significant impacts on whether a data is personal or not.

# How Data Environments Frame Whether Data is Personal

If one accepts the basic premise that the data is not personal or impersonal in isolation but only in relation to its context or environment, then it naturally follows that in some environments a given piece of data will be personal and others it will not. It further follows that together with the properties of the data itself, the elements that define the environment (all of which essentially control the likelihood of other data linking to the data in question) form the parameters through which a judgement can in principle be made as to whether data is personal or not.

Here we will consider three types of environment: (i) the open data environment; (ii) a community of users (created by licensing); and (iii) controlled safe settings. We start with the open data as that is the simplest to parameterise and then discuss the variation arising from the other two types of environment. The fundamental point is: however low a risk threshold one deems sufficiently low to make data non-personal (=functionally anonymous), whether data meets that threshold will depend on its environment as much the data itself.

## Open data

When data is made open it is in effect released into the global data environment. Any agent can in principle access the data and process it. Because in most jurisdictions open data is necessarily presumed to be non-personal data, there are no a priori restrictions on its processing.

An agent *Ui* has simultaneous access to four forms of auxiliary information (*I*) which could be used to re-identify population units within an open dataset *D*:

1. Data held within partitions to which *Ui* has access (*IP*).
2. Information available to *Ui* as personal knowledge – either their own or that of a third party (*IK*).
3. Other open data (*IO*).
4. Readily available data (*IR*).

As we shall see each of these forms has different properties.

As there are no governance constraints on *Ui*, *IP*+*IK*+*IO*+*IR* form the theoretical total data environment (*TEi*) for *Ui*. The attributes *Ai* (skills, metadata knowledge, resources and partition constraints on *Ui*) act as a functional constraint on that theoretical environment:



For the purposes of re-identification we consider the data environment with respect to both *Ui* and *D*:



Where ∩’ signifies that the intersection of *D* and *Ai(TEi)* is probabilistic due to the data divergence between the data and its environment. This function relates directly to the probability of an identification given attempt denoted by Marsh et al. (1991) in the equation:



That is, we are saying that the probability that an attempt will be successful given that it has taken place is captured by the attributes of the user in combination with the empirical intersection between the data that is being attacked and that data’s environment, and so:



Thus far we are only describing the risk of re-identification presuming that the user’s motivation to attempt that re-identification exceeds some disutility function over the costs of them doing so. So if we want to understand the likelihood of an attempt being made we need to take account of the agent’s level of motivation to attempt to re-identify somebody. There are many different models of motivation but discussion of these is beyond the scope of this article. Here, we will assume simply that there is some indicator function *M*, which for a given user will indicate that they will attempt to re-identify somebody in the data. That indicator function will itself be dependent on the data (*D*), the user’s functional attributes (*Ai*) and indeed *Ai(TEi) ∩’ D (*we assume that the user’s likelihood of attempting a re-identification will be affected by their perceived likelihood of success) and some utility (*U*) that the user obtains from the re-identification.

So then:



Finally, we can generalise this across all agents. The function *E(I)* is the expected number of identifications and *N* is number of agents with access to the open environment.



We have not here specified the indicator function *M* and given the aforesaid complexity any attempt to do so would be pre-theoretical. However, it is probably reasonable to assume that *M* is conceptually monotonic with respect to its parameters, so that a functional increase in any of those would lead to an increase in the probability that *M*=1.

## Licensing as community of trusted users

Licensing is primarily a form of soft environmental partitioning. Minimally, licensees undertake not to pass on the data to an unlicensed third party. Often there are an additional set of prescriptions (typically data security expectations) and proscriptions (typically an undertaking not to attempt re-identification of a data unit).

What impact does this have on the formulation given for the open data environment? Firstly, the number of agents is undoubtedly smaller. If the licensing was 100% functional then the number of agents would equal the number of licensed users. On more realistic assumptions – where we allow for some “leakage” – the number of agents is still considerably smaller than with open data. Against this it is reasonable to suppose the vast majority of those agents in the open data situation will have low values for *U* and *A*, and that this tail will be a large proportion of what is excluded by the licensing. Nevertheless, it is necessarily so that *N* in the licensing situation is a (small) subset of *N* in the open data situation.

A second difference is in the governance provided by the license. If we assume that there will be some consequence for the user for visibly breaking the license conditions then this will impact on *Ui* with the effect of decreasing the number of cases where *Mi*=1.

Even for those who have obtained a copy of the data outside of the license, we can reasonably assume that they have “behaved badly” and perhaps even illegally (e.g. hacking to obtain a copy of the data in contravention of the computer misuse act). This is a different situation to that of open data where the license basically allows anyone to pretty much do what they like with the data and specifically does not prohibit re-identification. One certainly could not generalise across all agents to say that this would always reduce U and indeed it is reasonable to assume that for some agents the reverse will be true that being “allowed” to do something will decrease the incentive to do so. However, even assuming that the mean effect on U was zero, the skills and resources required to carry out a re-identification will be greater. Furthermore, some partition constraints will be more significant; whilst I may even have legitimate reasons to link my company’s client database to an open data set with a licensed dataset that I have no legitimate access to, this becomes more problematic for my company’s reputation.

In short, the environmental elements of the right hand side of equation 1.6 are almost certainly smaller in the case of licensed data than in the case of open data. Therefore, it is reasonable to posit that the expected number of identifications is also lower. The implication of this is that – for a given level of *E(I)* – it is possible to release data such that  is larger with licensed data than with open data and that implies that *D* is either more detailed or less perturbed.

## Secure safe settings

Secure safe settings are a broad class of hard environmental partitions that use both soft and hard infrastructure and governance to control of access (who, why, when, where, how). UK Examples are the HMRC data lab and the Secure Data Service. One of the key differences between most forms of safe setting and licensed data is the hard partitioning. This directly limits *Ai*(*TEi)* in a way that licensing on its own does not. Common practice is to prohibit users from brining other data into the secure setting and strictly controlling the analytical output which can be taken out. This severely restricts the usefulness of *IP, IO, and IR.*

In effect an agent that wished to carry out a re-identification would be relying on *IK.* It is possible to imagine a user memorising information for a particular individual and then hunting for an individual with those attributes in the dataset but the wide-scale cross-match type attacks described by Elliot and Dale (1999) are ruled out. Indeed in some safe settings the user cannot even see the data so such attacks are ruled out.

The license conditions associated with the use of safe settings tend to be much heavier and so the consequences of carrying out a re-identification are more severe and the probability of being able to do so undetected is also much smaller.

So compared with simple licensing, safe settings have lower levels of both *Ai* (and therefore *Ai(TEi))* and *Ui.*and so consequently lower levels of *E(I),* given a fixed dataset. How much lower will depend on the details of the governance and infrastructure of the safe setting. Virtual safe settings (such as the Australian RADL or the UK’s Secure Data Service) will have higher levels of *E(I)* compared to on-site labs; settings where the user is able to view the data directly will have higher levels than those where there is no such function (for example the English Census Longitudinal Study).

# Concluding Remarks

In this article we have addressed the question of the extent to which the nature of the environment within which data are held (their *data environment*) can make data (which would otherwise be deemed personal) non-personal. In our discussion of anonymisation we have described how that data which is *absolutely anonymised* is unlikely to be useful and that data which has only been *formally anonymised* is likely to still be personal.

We have developed the concept of *functional anonymisation* that ties together notions of disclosure risk with those of the data environment. We have proposed that a data environment can be understood by a small number of parameters: other data in the environment, the skills, knowledge and motivations of the persons who are present in the environment, governance structures and process, and infrastructure and we have argued that those parameters are to a large extent controllable.

In essence then, functional anonymisation is the practice of reducing the risk of re-identification (through controls on the data and its environment) so that it is at a sufficiently low level that that data in that environment can be deemed to be non-personal. This leaves open the question of how small a value of *E(I)* would be deemed be sufficiently low for us to regard a dataset as non-personal. This is ultimately a policy decision that is outside the scope of this article. However, we note that these sorts of decisions are made all the time in human societies (e.g. what is the expected number of serious radiation leaks from nuclear power stations that is deemed to be sufficiently low so that we can call the power station programme safe?). The reason for these policy decisions is because we want the social goods that are the upside of the decision (available electricity) and so it is with personal and non personal data.

The overall point here is that the potential benefits of the use of data about us for our society and for us as individuals our huge. However, privacy must not be thrown away in the process of striving for those benefits. Functional anonymisation is a practical framework for delivering the desired benefits whilst protecting individual privacy.

# References

Bonnici, J. P. M. (2014) “Exploring the non-absolute nature of the right to data protection”. *International Review of Law, Computers & Technology*, 28:2, 131-143, DOI:

Duncan, G., Elliot, M., and Salazar-González, J. J. (2011) *Statistical confidentiality: principles and practice*. New York: Springer.

Dwork, C. (2006) “Differential privacy” In *Automata, languages and programming* 1-12. Berlin Heidelberg: Springer.

Elliot, M., and Dale, A. (1999) “Scenarios of attack: the data intruder’s perspective on statistical disclosure risk” *Netherlands Official Statistics*, 14(Spring): 6-10.

Elliot, M., Lomax, S., Mackey, E., and Purdam, K. (2011) “Data environment analysis and the key variable mapping system” In *Privacy in Statistical Databases* 138-147. Berlin Heidelberg: Springer.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P. P. (2012) *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons.

Koops, B.J. and Leenes, R. (2014) “Privacy regulation cannot be hardcoded. A critical comment on the ‘privacy by design’ provision in data-protection law”. *International Review of Law, Computers & Technology*, Volume 28, Issue 2, 159-171.

Mackey, E. C., and Elliot, M. J. (2009) “An application of game theory to understanding statistical disclosure events” In proceedings of *Joint UNECE/Eurostat work session on statistical data confidentiality* (Bilbao, Spain, 2-4 December 2009), available online at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.40.e.pdf

Mackey, E and Elliot, M. J. (2013) “Understanding the Data Environment.” *XRDS: Crossroads*, 20 (1): 36-39.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991): "The case for samples of anonymized records from the 1991 census." *Journal of the Royal Statistical Society* *series A*, 154: 305-340.

Muralidhar, K., and Sarathy, R. (2011) “Does Differential Privacy Protect Terry Gross’ Privacy?” In *Privacy in Statistical Databases* (pp. 200-209). Berlin Heidelberg: Springer.

Ohm, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. Rev. 1701

Paass, G. (1988) "Disclosure risk and disclosure avoidance for microdata", *Journal of Business and Economic Statistics,* 6(4): 487-500.

Wright, D. and Raabb, C. (2014) “Privacy principles, risks and harms.” *International Review of Law, Computers & Technology,* Volume 28, Issue 3, 277-298.

1. For example, the UK ICO ‘Anonymisation: managing data protection risk code of practice’. The Code demonstrates that pseudonymous data can be transformed into anonymous non-personal data where suitable privacy enhancements are in place. It is aimed at organisations wanting to transform personal data into anonymised information for use in research or for other data analysis purposes. The Code explains the issues surrounding the anonymisation of personal data, and the disclosure of data once it has been anonymised. [↑](#endnote-ref-1)
2. This can also impact on the operation of other legislation for example if data is non-personal in the recipient data environment then FOI requests can be acted upon. [↑](#endnote-ref-2)
3. Perhaps most important is the freedom to disclose the data. Disclosure here means allowing access to the data by one or more third parties which could come in a variety forms: publishing, disseminating, sharing, licensed access, controlled access etc. [↑](#endnote-ref-3)
4. It should be noted here that disclosure control researchers distinguish between *identification* and *attribution* processes in a disclosure. The former indicates that agent X has found person Y in some (supposedly anonymised) data, the later indicates that agent X has learnt something new about person Y. These two processes often co-occur but need not. This is somewhat confusing because the two processes are conflated in data protection law; thus in the *Anonymisation code of practice* the UK Information Commissioner says “Note that ‘identified’ does not necessarily mean ‘named’. It can be enough to be able to establish a reliable connection between particular data and a known individual.”; page 21. [↑](#endnote-ref-4)
5. Also, without users there would be no practice of symbolism, and therefore no data. [↑](#endnote-ref-5)
6. http://www.legislation.gov.uk/ukpga/1998/29/section/1 [↑](#endnote-ref-6)