OPEN FOR FEEDBACK

WORK IN PROGRESS

Anonymisation and open data

An introduction to managing the risk of re-identification

Open Data Institute

Table of contents

Anonymisation and open data

An introduction to managing the risk of re-identification

About

This report has been researched and produced by the Open Data Institute (ODI), and published in April 2019. Its lead authors are Olivier Thereaux and Fionntán O'Donnell with contributions from Sonia Duarte, Becky Ghani, Jared Robert Keller, Jeni Tennison and Peter Wells.

The report follows the recent implementation of the EU's General Data Protection Regulation (GDPR) in May 2018. It is one of the most comprehensive pieces of regulation yet to deal with personal data, and is changing the way organisations collect and use data.

GDPR is just one part of a growing trend towards people exercising more control over data about them, and protecting their privacy in a data-hungry world.

We hope that this report will provide organisations with a better understanding of data anonymisation as a framework for risk management, and how responsible data management can increase access to data while retaining trust.

We would like to thank Mark Elliott and Elaine Mackey from The University of Manchester, Gemma Galdon Clavell and Carmela Troncoso from Eticas Research and Consulting, and Yves-Alexandre de Montjoye from Imperial College for their expertise, time and advice to our team throughout our research. We would also like to thank Jonathan Pearson and Forrest Frankovitch at NHS England, as well as the team at ODI Leeds, for providing us with a practical case to test our tools and understanding.

April 2019

If you would like to send us feedback, please get in touch by email at RandD@theodi.org.

Open data is (sometimes) made of people

Open data initiatives have historically been motivated by increasing transparency about governments and support for economic growth[1]. Typical open data describes things like bus stop locations, the weather forecast, or local authority payments. Open data, from a distance, is about things and organisations, not people.

DEFINITION

Open data is data that is available for anyone to access, use, and share.

The ODI's Data Spectrum[2] illustrates how data access and governance sits on a spectrum of, from closed to shared to open. The current version of the spectrum includes personal data at the closed and shared end of the spectrum.

This may lead to a perceived dichotomy – that open data never includes personally identifiable information. Sometimes this perception extends to thinking that data derived from personal data – i.e. anonymised, pseudonymised or synthesised data – cannot be open data.

This perception was apparent in the results of some user research we conducted in late 2018[3]. The interviewees had a very distinct mental model: on the one hand, personal data which is meant to be kept closed and protected; on the other, open data, which almost never includes personal data.

The reality, however, is more subtle than this perception suggests.

For example, the expenses and allowances for elected officials in the UK, such as council members[4] or Members of Parliament (MPs)[5], are open data.

There are also many examples of open datasets for licensed practitioners, eg the General Medical Council's list of registered medical practitioners[6]. Which professions are included varies from country to country, from legislation to legislation. For example, as part of the *Loi pour une république numérique*[7] (digital republic law), France has recently enshrined into law[8] which personal data ought to be published as open data, for the public good.

---

[1] The Guardian (2013), *Obama to Berners-Lee, Snow to Domesday: a history of open data,* https://www.theguardian.com/news/datablog/2013/oct/25/barack-obama-tim-berners-lee-open-data

[2] Open Data Institute (2016) *The Data Spectrum*, https://theodi.org/about-the-odi/the-data-spectrum/

[3] Open Data Institute (2019) *How do organisations perceive the risks of re-identification?*, https://theodi.org/article/how-do-organisations-perceive-the-risks-of-re-identification/

[4] Lincolnshire County Council (2017) *Members' allowances,* https://data.gov.uk/dataset/3b17b920-642a-4189-b1d9-18e420be9ef4/members-allowances

[5] *Independent Parliamentary Standards Authority,* https://www.theipsa.org.uk

[6] General Medical Council (2019), 'List of registered medical practitioners', https://webcache.gmc-uk.org/gmclrmp_enu/start.swe

[7] Legifrance.gouv.fr (2016), 'Loi pour une république numérique', https://www.legifrance.gouv.fr/affichLoiPubliee.do?idDocument=JORFDOLE000031589829&type=general&legislature=14

[8] Etalab (2018), *Le décret fixant les catégories de données diffusables et réutilisables sans anonymisation est paru,* https://www.etalab.gouv.fr/le-decret-fixant-les-categories-de-donnees-diffusables-et-reutilisables-sans-anonymisation-est-paru

In both cases – public office transparency and legal requirements – the open data includes the names of officials or professionals, but also potentially reveals information about their life or their habits. This is deemed an acceptable encroachment of their right to privacy, because that right is balanced against other rights, such as the right of the public to hold those in public office to account[9], or to guard against fraud or unlicensed practice.

Personal, sensitive and private data

There may be value in opening data about people, but this must be balanced with the risk of harm.

In some instances, the case to make data public and open overrides concerns because there are competing rights between the people who are potentially impacted by its availability and use. Sometimes societies decide that to protect people from harm we need to either risk, or even cause, harm to some people. At the simplest level this is why nations have the power to punish people or put them in prison. Different societies make their own decisions about which crimes receive which punishment. For example, the USA publishes its sex offender registry[10] as open data with the public interest goals of reducing instances of sexual violence and re-offending[11].

In many other cases, open access to personal data could bring benefits, for example for researchers or innovators, but the potential harm to people outweigh those benefits. But is there a way to protect privacy while still realising the value associated with open data?

The notions of personal and sensitive data are not just intuitive: they are reflected in official definitions used in laws and regulations.

DEFINITIONS

The EU General Data Protection Regulation[12] (GDPR) defines ***personal data*** as:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

The UK's Office for National Statistics defines ***private information***[13] as*:*

> information that
>
> - relates to an identifiable legal or natural person, and
> - is not in the public domain or common knowledge, and
> - if disclosed would cause them damage, harm or distress

[9] Committee on Standards in Public Life (2016), 'Upholding the Seven Principles of Public Life in Regulation Report' https://www.gov.uk/government/publications/striking-the-balance-upholding-the-7-principles-in-regulation

[10] DC Court Services and Offender Supervision Agency (2015) Sex Offender Registry, http://opendata.dc.gov/datasets/10e58174831e49a2aebaa129cc1c3bd5_20

[11] National Criminal Justice Reference Service (2011), 'Evaluating the effectiveness of sex offender registration', https://www.ncjrs.gov/pdffiles1/nij/grants/234598.pdf

[12] European Union (2016), Regulation (EU) 2016/679 (General Data Protection Regulation) Art 4. Definitions, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[13] Office for National Statistics (2009), *National Statistician's Guidance on Confidentiality of Official Statistics*, https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Confidentiality-of-Official-Statistics-National-Statisticians-Guidance.pdf

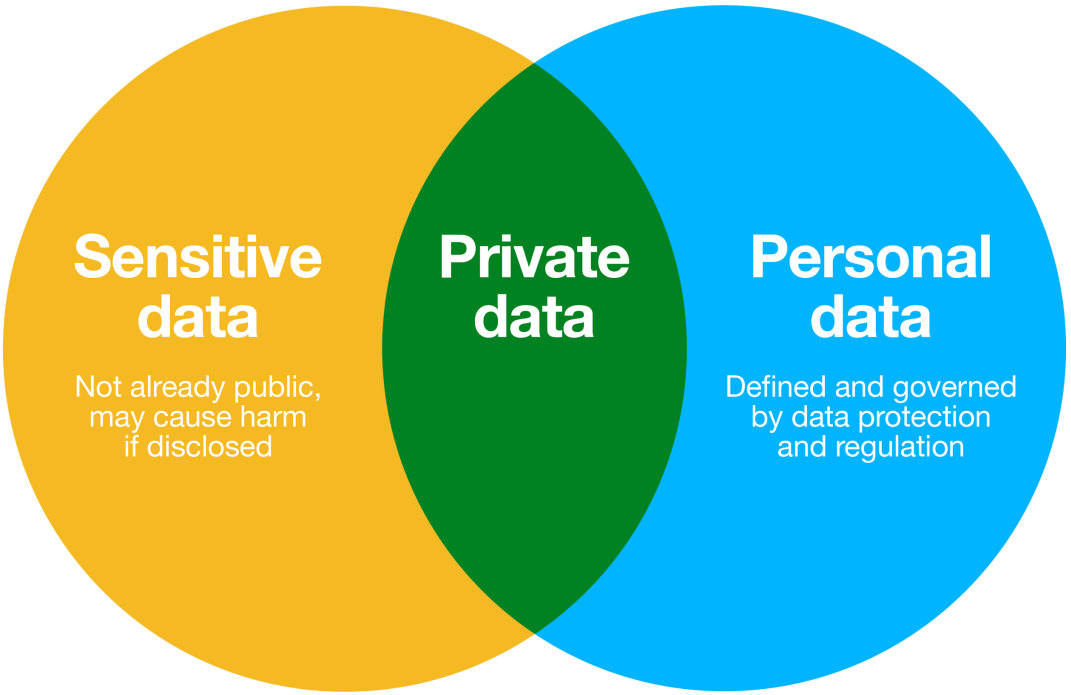
This definition is very similar to what GDPR calls ***Special category data***[14]

Black's Law Dictionary defines ***sensitive information***[15] as:

> A sensitive asset that if compromised can cause serious harm to an organisation

Personal data can be sensitive data – in which case, in line with the definition above, we call it private data. Or, in the case of a list of elected officials, personal data can be deemed non-sensitive. Conversely, sensitive data can be personal data, but it does not have to be: corporate confidential information[16] or environmental protection data (like the locations of endangered species, or animals protected from poaching) are all examples of sensitive, but not necessarily personal, data.

We can visualise this overlap between sensitive, personal and private data as:



How does open data relate to this diagram?

Whether data is personal is inherent to the data itself. Whether it is sensitive is dependent on both the content of the data and the wider information environment at a point in time. Whether data is made open data is a choice. The examples above show that personal data can often be made open – when in the public interest.

When data is made open, by definition it is no longer sensitive. The definition of sensitive data is that it is 'not already public'; open data is public.

---

[14] Information Commissioner's Office (2018), 'Guide to Data Protection' https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/

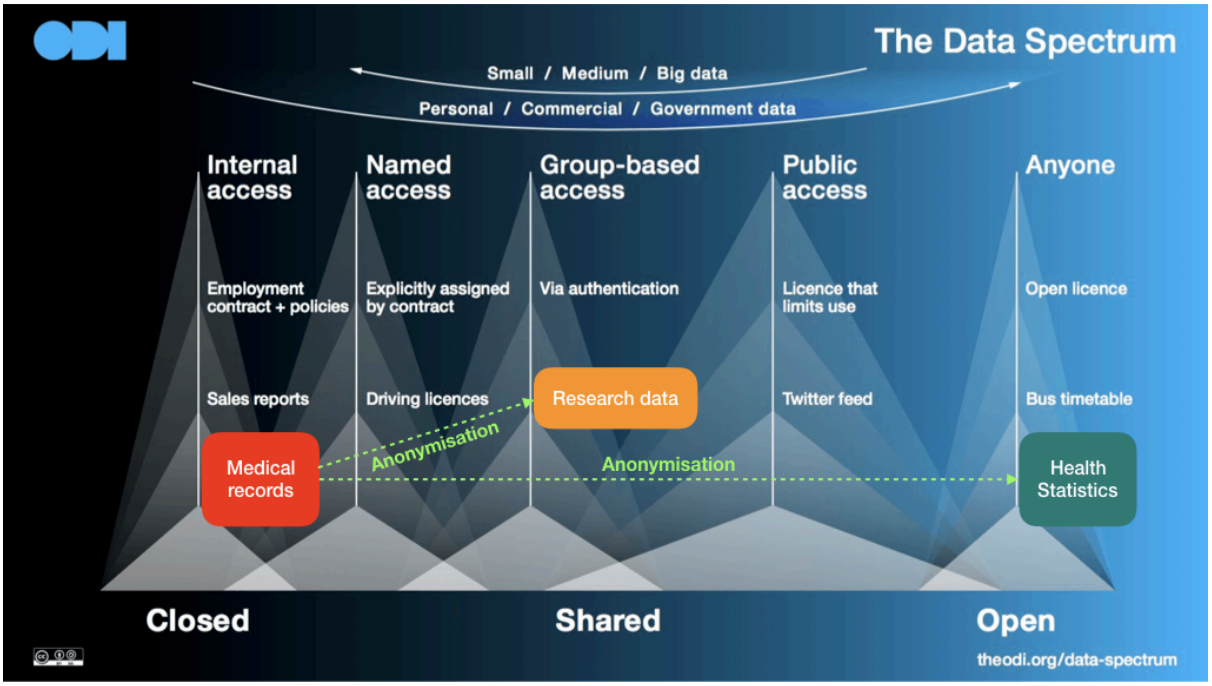
[15] *Black's Law Dictionary Free 2nd* Ed, 'What is sensitive information?', https://thelawdictionary.org/sensitive-information/

[16] Open Contracting Partnership (2019). Mythbusting: confidentiality in public contracting http://mythbusting.open-contracting.org

Redaction and anonymisation

It is however possible to process data into a modified form that can be shared or made open while significantly reducing (or eliminating) the possibility of anyone recovering sensitive or personal information from it. For sensitive data in general, this process is called redaction; for personal data it is called anonymisation.

For example, we can visualise how anonymisation of medical records would create modified forms of data that could sit elsewhere in the Data Spectrum – shared/open rather than closed. This would allow research data to be shared with academics through data-sharing agreements, and health statistics to be published openly:



How common is anonymised/personal open data?

So far we have only partly challenged the perception that personal data and open data are mutually exclusive, by giving examples of open personal data and introducing the concept of anonymised open data. Are those, however, anecdotal occurrences or more commonplace?

In February 2019, our team conducted a small-scale audit of open data portals[17], taking a random sample of 114 datasets published to the French[18] and UK government open data portals[19], and assessing whether each of the selected datasets contained personal information, had been anonymised, or both.

We found that one in five datasets sampled on the French portal, and one in ten on the UK portal, included some personal data – often names of people, whether because they were elected officials, registered professionals, or, in the case of the French portal, because the dataset listed sole traders or companies named after their owner, thus including personally identifiable information.

The proportion of datasets including anonymised data was around one in five in both cases, with one caveat: qualifying a dataset as 'anonymised' can be approximative and uncertain without information about how it was processed. We would need a full

---

[17] *Open Data Institute (2019), 'Auditing data portals for personal or anonymised data',* *https://docs.google.com/spreadsheets/d/1fkmV5k8FANXXkLclu7pgZQQrxOgWeCsazXafr_EMUYE/edit#gid=0*

[18] data.gouv.fr, https://www.data.gouv.fr/

[19] data.gov.uk, https://data.gov.uk/

description of the data processing that led to the open form of the data to fully verify its status as anonymised.

Some cases were relatively easy to gauge, such as when a dataset included a 'name' column with values removed, or when the description of the data noted how it had been anonymised. In other cases, we considered personal data to be anonymised when aggregated or otherwise obfuscated so that no personally identifiable information remained.

Anonymisation, utility and risk

There are several, sometimes contradictory, definitions of anonymisation and related terms such as de-identification, and how they relate to one another.

DEFINITIONS

For the rest of this document, we will refer to the definitions used in the literature review commissioned by the ODI and created by Eticas, a research and consultancy company with expertise in security and responsible technology.

Anonymisation:
A process that alters a dataset to reduce the risk of re-identification as much as possible.

Adversary:
An entity with access to an anonymised dataset that seeks re-identification of an individual or to learn more information about them

Note that in this definition, the adversary does not necessarily intend to cause harm. Therefore the anonymisation processes described also cover the prevention of accidental disclosure or discovery of private information.

Not just a technical tool

It is important to note that here, anonymisation is defined as a process – a series of actions. Not just a technique, not just a tool, and not just a set of mathematical equations.

Data practitioners may be tempted to approach anonymisation as a purely or predominantly technical activity. Once a team has chosen anonymisation as the best way to safely gather, process or share data, choosing an appropriate statistical disclosure method may appear to be the logical next step. However, anonymisation is a more in-depth process than this, involving research, legal and ethical considerations, risk analysis and testing.

In a guide first published in 2016 (and to be updated in 2019), the UK Anonymisation Network (UKAN) offered a practical and integrated approach: the Anonymisation Decision-Making Framework[20] (ADF).

One of the main messages of the ADF is: anonymisation is not done in a vacuum. To anonymise effectively, the process should include serious thinking about what happens to the data after anonymisation; with whom will it be shared; what could go wrong; and what mitigations can be put in place.

The right tools and techniques can only be employed effectively by going through this process of understanding the data ecosystem and data flow; engaging with internal

[20] UKAN (2019), *The anonymisation decision-making framework,* *https://ukanon.net/ukan-resources/ukan-decision-making-framework/*

and external stakeholders; understanding the legal and ethical context; and thoroughly considering the risks and their mitigations.

There is more than one way to do it

Anonymisation is sometimes misunderstood as the act of simply removing personal identifiers (mostly, people's names) from data sets. This technique, often known as 'de-identification' may be appropriate in some cases, but it is a very limited technique: removing only explicit identifiers often leaves information within the dataset that can still be used to identify an individual.

The field of anonymisation has however greatly evolved in the past three or four decades. Contemporary anonymisation techniques are now many and varied, generally falling into three categories:

1. **Suppression:** removing identifiers or pieces of information that may lead to re-identification
2. **Generalisation:** aggregating data points into a coarser granularity, or otherwise removing details to obfuscate data about people on an individual basis
3. **Disruption:** adding noise and changing values to the extent that it is increasingly difficult to know how, or whether, information about specific individuals can be recovered or inferred.

Not just how, but how much

This document will not attempt to give a detailed explanation of the variety of the techniques – but we would recommend the documents referred to throughout this report and listed in Appendix 1.

We will, however, consider a simple case of anonymisation to illustrate how the techniques can be applied.

Example: civil servants' crumpet-eating habits

In this example, an organisation ran a survey of all civil servants in the UK, collecting data on how many crumpets the civil servants typically eat in one sitting. The organisation was commissioned to help the civil service optimise its crumpet purchase, but is now considering whether to release an anonymised version of the survey results for the benefit of the crumpets-and-other-foodstuffs-industry.

The raw data could look a little bit like this:

Full name	Date of birth	Typical number of crumpets eaten in one sitting
Peter Watson	1969-03-16	5
Lydia Okwanga	1985-12-02	1
...	...	...
Ali Massa	2000-05-26	2

The data steward (who collects, maintains and shares data)[21] in charge of releasing an anonymised version of this data may decide that the first technical step is to remove all names (*suppression*), replacing them with pseudo-identifiers (*pseudonymisation* – a type of *suppression*):

ID	Date of birth	Typical number of crumpets eaten in one sitting
1	1969-03-16	5
2	1985-12-02	1
...	...	...
35769	2000-05-26	2

While removing names is an obvious step, if you know a colleague's birthday you could easily identify them in the data, and infer how many crumpets they ate. The data steward may therefore decide to also replace the very precise dates of birth with age ranges (*generalisation*):

ID	Age range, as of 2019-02	Typical number of crumpets eaten in one sitting
1	40–50	5
2	30–40	1
...	...	...
35769	10–20	2

Crumpet-eating habits are quite sensitive, so they may also want to add some "noise" to the data, for example by swapping values for a certain proportion of the rows (disruption):

ID	Age range, as of 2019-02	Typical number of crumpets eaten in one sitting (up to 10% of the values have been swapped between rows)
1	40–50	1
2	30–40	5
...	...	...
35769	10–20	2

[21] Note on language: In this document we use the term 'data steward' to describe a person or organisation who collects, maintains and shares data. We may review the choice of wording in future iterations. Under GDPR, this role is defined as a 'data controller' or 'data processor'. For GDPR definitions, see https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/

Each of those iterations results in a form of 'anonymised data' – as would an otherwise infinite combination of anonymisation techniques.

With each of those steps, the data steward removed a certain amount of detail, granularity and precision from the data. This means the data is less specific and less accurate. Anonymisation reduces its 'utility' – a measurement of the usefulness of the data.

Utility is a relative concept: after the first two anonymisation steps above, the dataset will be less useful to someone needing to answer the question: "how many crumpets does Peter typically eat?", but utility would still be high for someone trying to answer a question about whether, in this dataset, people in their 20s eat more or fewer crumpets than people in their 50s.

Utility is typically at its maximum in the raw data because that data supports the widest possible range of uses, although there too it is relative: the full, non-anonymised raw data will for example be more useful to someone with access to data about where Peter, Lydia and Ali live. At the other end of the spectrum, utility will generally be considered to reach a minimum if anonymisation is so heavy that it renders the data useless, as in this (extreme) example:

Typical number of crumpets eaten in one sitting (anonymised)
Some people eat crumpets some of the time.

The utility-risk trade-off

Data stewards, whether they hold data about crumpet consumption or other topics, have at their disposal an almost infinite combination of anonymisation techniques and parameters. They could apply different forms of anonymisation to the same raw data to produce several different anonymised datasets, and release or share all of them. That's a large amount of what census-based statistics do: they are different aggregations, focusing on different features, a bit like showing only parts of the elephant in the parable of the six blind men.

Too much choice can however be paralysing. Assuming they want to release a single anonymised dataset about crumpet consumption, how should our data steward decide how much to anonymise this data, and how much utility should remain?

To answer this question, we need to get back to why they would go through a process of anonymisation in the first place: to address the risk of accidental disclosure of sensitive information – or of re-identification by an adversary, both of which could cause harm to the data subjects.

Risk management is a whole discipline dealing with the assessment of the probability and impact of risk, and its mitigation. This document does not describe the discipline at length, but resources listed in Appendix 1 address how to model threats, how to gauge the likelihood and impact of those threats, and how to manage associated risk. The ADF, in particular, offers guidance on 'building disclosure scenarios'.

What is important to understand, however, is the correlation between utility and risk in the anonymisation process. In *Broken Promises of Privacy*[22], professor Paul Ohm writes that: '*Data can be either useful or perfectly anonymous but never both*'.

[22] Paul Ohm (2010), 'Broken promises of privacy: Responding to the surprising failure of anonymization'.
https://paulohm.com/classes/infopriv13/files/week8/ExcerptOhmBrokenPromises.pdf

It would be tempting, but a mistake, to extrapolate from this statement that, since perfect anonymisation is impossible, anonymisation is pointless. Ohm explains further: '*No useful database can ever be perfectly anonymous, and as the utility of data increases, the privacy decreases*'. In other words, every decision in anonymisation is a trade-off between risk and utility.

Open data and re-identification risk

This is where we finally answer the question posed earlier: why isn't anonymised open data more common?

We now understand that decisions about anonymisation of data are decisions about risk and utility – alongside compliance with data protection laws – based on an assessment of threats in a specific context, such as who will the data be shared with, when, and how.

Open data is one such context, and it is one of the most complex:

1. With open data, most of the factors used in assessing, planning, conducting and monitoring the anonymisation process are unknown by design. Open data is available for anyone to access, use and share, sometimes in unanticipated ways. The adversary that needs to be considered may therefore be anybody, at any point in the future[23]. They may also have access to any data and any current or future technologies.

2. Risk is therefore hard to assess. The process of risk assessment in anonymisation presumes that – through analysis and imagination – the data steward is able to make assumptions about who the adversaries may be, what other data might be at their disposal, when and for how long the adversary may attempt de-anonymisation, etc. Those assumptions cannot apply in the case of open data.

Utility is also harder to assess. Guides like the ADF involve a consideration of how the data will be shared and used: it helps data stewards understand how 'useful' the data will be post processing. But without full knowledge of possible uses – common in the open data situation where serendipitous reuse is a feature – the understanding of utility is imperfect, and therefore anonymisation decisions are less certain.

Another perspective is to look at the Data Spectrum as a spectrum of governance and certainty.

- The closed end of the spectrum has the most certainty about who may access and use the data, and the strictest governance in terms of access, rights, etc.
- In the shared space there are still rules and governance – regardless of which sharing model is chosen – and some clarity over who may access and use the data and when, but there is some uncertainty, and therefore some risk, which can then be assessed and mitigated through anonymisation.
- The open end of the Data Spectrum creates the most potential for access, use and processing, but the more permissive the governance, the more uncertainty it creates.

Lies, damned lies, and statistics

The risk of re-identification is real. In a set of case studies[24] commissioned by the ODI, Eticas described a number of high-profile cases where anonymisation, albeit well thought through, resulted in re-identification or disclosure.

---

[23] The life span of anonymised data is typically estimated at 100 years - after which most if not all of the individual data subjects are no longer likely to be alive, therefore mooting the risk of re-identifying a living person.

[24] Eticas Consulting (2019), 'Responsible Open Data' https://eticasfoundation.org/wp-content/uploads/2020/05/ODI-D2-final.pdf

It is worth noting that many cases of re-identification or 'breaking' of anonymisation are not necessarily a clever mathematical anonymisation technique being defeated by a cleverer mathematical de-anonymisation technique, but simply a result of inadequate or clumsy anonymisation.

This excerpt from a high-profile 2018 court case[25] provides a rather comical example of bad anonymisation (in this case suppression through pseudonymisation), literally pointing to the single and very publicly available external data point required for re-identification, namely: *who was president of the United States in January 2017?"*

> board of directors, thereby fixing the problem for Individual-1. (*Id.*) Not long after, Cohen was hired by the Company to the position of "Executive Vice President" and "Special Counsel" to Individual-1. (*Id.*) He earned approximately $500,000 per year in that position. (*Id.*)
>
> In January 2017, Cohen formally left the Company and began holding himself out as the "personal attorney" to Individual-1, who at that point had become the President of the United States. In January 2017, Cohen also launched two companies: Michael D. Cohen and Associates,

There is always a probability that an adversary with strong motivation and high skills and resources will be successful in attacking anonymised data. That said, poorly anonymised data – due to lack of skills or scrutiny – is where a significant amount of risk lies.

In our exploration of open data portals, we found at least one example of a dataset which had gone through a process of anonymisation, but included people's names in a 'notes' field which had not been removed.

Conversely, while anonymised data remains a minority of the data released under open licence, and most anonymised data is typically shared under much stricter governance, there are plenty of examples of data being anonymised and published openly, while retaining significant utility.

The release of statistics is a major use case for 'anonymised open data done right'.

Statisticians routinely release open data, with risk reduced to a point that few would challenge their release for fear of privacy harm. Based on our conversations with members of the UK's Government Statistical Service Open Data sub-group[26], including members of the Data, Statistics and Digital Identity Division in the Scottish Government, this is mainly because of three things:

1. First, statistics professionals have a deep understanding of statistical disclosure controls: methods for ensuring aggregate statistics do not inadvertently reveal information about individuals.

2. Second, statistics bodies are keenly aware of the balance between privacy and other rights and outcomes. In its Guiding Principles for Data Linkage[27], the Scottish government clearly aims for the balance between risk and utility we

[25] United States District Court, southern district of New York (2018), 'Sentencing Memorandum, UNITED STATES OF AMERICA v MICHAEL COHEN'
[26] Government Statistical Service (2018), Open Data, https://gss.civilservice.gov.uk/guidances/open-data/
[27] Scottish Government (2012), 'Joined-up data for better decisions: Guiding Principles for Data Linkage', https://www.gov.scot/publications/joined-up-data-better-decisions-guiding-principles-data-linkage/

mentioned earlier:

"*The law does not give absolute value to privacy and therefore a balance is needed between respect for privacy, through the proportionate mitigation of risk, and the potential benefits to all through the use of data for statistical and research purposes.*"

It also needs to balance other considerations, such as the cost and complexity of the process: "It is not in the public interest to undertake unnecessary anonymisation work if the risk is low, as it increases public spend with very little or no benefit."

3. Finally, one key reason why statistical bodies can routinely publish anonymised open data is that they are dealing with such volumes of data that even with a big drop in utility, what they release through aggregation methods is still extremely useful.

This way of thinking is not limited to statistics from public bodies: in a very similar setting, corporations publish open data[28] about their operations for transparency purposes, but aggregated at such a level that any sensitive information is not visible.

What the future holds

While traditional anonymisation techniques have been around for decades, the age of large-scale data processing and machine learning has brought two important innovations to the domain.

Differential privacy

Over the last 15 years, differential privacy has gone from theoretical academic research to being deployed by software multinationals in smartphones and browsers to protect the privacy of hundreds of millions of users.

Differential privacy is a property of data systems that allows collection of aggregated statistics about a dataset but obfuscates individual records. When queried, a small amount of noise is added to the data such that if any one record were removed, the query result would stay the same. This means those using the data can never be entirely certain about any single person's data.

The first large commercial use of differential privacy was in 2014, with Google trialing it in the world's most popular web browser, Chrome, to collect statistics on malware[29]. In 2016 Apple added differential privacy to its *iOS 10* operating system to collect information on keystroke usage and browser performance[30]. Uber has also published a tool it uses to study patterns of car rides without accessing passenger's location information[31].

While obviously powerful, differential privacy should not be seen as the answer to all data anonymisation problems. It works best in situations where one organisation has

---

[28] Syngenta (2018), 'Open data agriculture' https://www.syngenta.com/who-we-are/our-stories/open-data-agriculture

[29] Google (2014), 'Learning statistics with privacy, aided by the flip of a coin', https://security.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html

[30] Apple (2017), 'Differential Privacy', https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

[31] Uber/Github (2016), 'Dataflow analysis & differential privacy for SQL queries' *sql-differential-privacy* toolkit on Github. https://github.com/uber/sql-differential-privacy

sole access to a large dataset collected from lots of people. It requires those who use it to predict the number of queries that will be made on a dataset, to know how much noise to add. Also, it is mainly focused on simple data queries, where the answer is a binary choice or a single number. For more complex data, differential privacy could add too much noise and render the results useless[32].

Generating synthetic data with deep-learning algorithms

Synthetic data is data that is created by an automated process such that it holds similar statistical patterns as an original dataset. Each individual record in the synthetic dataset may have no relation to reality but when viewed in aggregate the dataset is still useful for certain analyses and for testing software. If done correctly, the synthetic data can contain no personal data even though it is based on a dataset that holds personal data.

Academics have researched the use of synthetic data since the early 1990s[33] using standard data-analysis and machine-learning algorithms like parametric modelling, linear regression, and bayesian networks.

However, in the past few years, a sub-branch of machine learning called deep learning, which uses large datasets and neural networks, has made huge advancements in the generation of synthetic data. Deep-learning algorithms have recently gained public fame and notoriety as they have been used to create fake audio, videos and images of famous people, places and, of course, cats[34]. They can be used to generate realistic-seeming numerical and natural-language data. That is, they can be adapted to create anonymised synthetic data from standard structured datasets.

Recently there has been an emergence of new companies specialising in synthetic data generated through deep-learning, with some specifically focusing on anonymisation. Their services include helping clients get past the 'privacy bottleneck' and share data without being blocked by the legal and ethical risks associated with processing personal data.

In the public sector, the UK's Office for National Statistics has done some early investigative work to see if synthetic data generated through deep-learning could be used to increase access to datasets[35]. It found promising results, saying the neural networks are *'excellent at approximating the data distribution of the original datasets, uncovering the underlying patterns in the datasets'*. However, they warned that the models can be *'susceptible to statistical noise'* and that synthetic data *'should always be mixed with reality to make informed decisions'*.

---

[32] Access Now (2017), 'Differential privacy, part 2: It's complicated.', https://www.accessnow.org/differential-privacy-part-2-complicated/

[33] Ioannis Kaloskampis, ONS Data Science Campus (2019), Synthetic data for public good, https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

[34] Jonathan Hui (2018), 'Some cool applications of GANs', https://medium.com/@jonathan_hui/gan-some-cool-applications-of-gans-4c9ecca35900

[35] Chaitanya Joshi, ONS Data Science Campus (2019), 'Generative adversarial networks (GANs) for synthetic dataset generation with binary classes.' https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/

Appendix 1: Going further

The goal of this 'primer' is to present an honest picture of why anonymisation for open data is hard, but not impossible, and to give an overview of the field of anonymisation.

There are many more resources available for the data practitioner – whether you are aiming to publish or share data through anonymisation, or using data which may have been anonymised.

Here we have listed resources which may help you go further, several of which were produced, or commissioned by the ODI in the course of our project on managing the risk of re-identification.

1. Anonymisation: a literature review – read more on the history and research on anonymisation
2. Anonymisation: case studies – Learn from examples of successful (and unsuccessful) anonymisation in this case studies document
3. Anonymisation Decision-Making Framework (ADF) – The UK Anonymisation Network offers a comprehensive guide to the anonymisation process
4. The ODI has created a prototype of a simplified step-by-step interactive guide to the Anonymisation Decision-Making Framework, using our eLearning platform
5. Anonymisation: A Short Guide – a high-level step-by-step guide to anonymisation, focusing on techniques rather than process, is also available as part of Eticas's work on responsible open data
6. Guide to the General Data Protection Regulation (GDPR) - provided by the Information Commissioner's Office (ICO) in the UK
7. Anonymisation: managing data protection risk code of practice – published by the ICO in 2015, provides practical advice on methods for anonymising data and the associated risks
8. Anonymisation: register of actors The ODI and Eticas have produced a list of actors in the field, from academia to the private sector, who can help with anonymisation.