Deliverable 4:

How to anonymise personal data

ODI Tender RDP8-001

Authors: Gemma Galdon Clavell and Carmela Troncoso

Research assistants: Victoria Peuvrelle and Miguel Vallbuena

Eticas Research and Consulting

C/ Ferlandina, 49

08001, Barcelona

(+34) 93 600 54 00

www.eticasconsulting.com

Table of contents

Introduction

There are several guides, frameworks and policies to help manage the release of data and add responsibility to open data processes. The United Kingdom's Information Commissioner's Office (ICO), for instance, has released an anonymisation code of practice aimed at any organisation, be it private, public or third sector, aiming to anonymise data.[1] Their code is not technical and does not provide in-depth information on matters of security engineering or statistical methodology. It is also not a step-by-step guide, and focuses instead on the management of the risks surrounding anonymisation. The UK Anonymisation Network, an organisation coordinated by a consortium of 4 British universities, also published general guidelines in their *Anonymisation Decision-Making Framework*.[2] In this document, they go over anonymisation concepts and define ten anonymisation components (steps) and three core activities (audit, risk analysis and impact management).

Focusing on a specific type of data, the European Medicines Agency published an external guidance on their 0070 policy,[3] where they define different anonymisation techniques, provide guidance on the process of anonymising clinical reports and include an anonymisation report template. The British Office for National Statistics, on the other hand, has a policy for social survey microdata release published on their website.[4] This guide is useful for data coming from sources such as census data and places a special emphasis on the different types of data release: open data vs. restricted or controlled access.

The following step-by-step guide takes into account the above-mentioned state of the art, but also complements existing frameworks by incorporating both technical and non-technical procedures (from techniques to governance models), aiming to create a succinct, practical document to help practitioners cover all the necessary steps to achieve robust pseudonymity.

A step-by-step guide to anonymising data

The previous sections introduced the use cases, exemplifying the difficulty of anonymisation with failures. They also provided an overview of the legal framework, and of techniques that have been used, or proposed, tailored to the use case in question. In this section we take a more hands-on approach, and we provide guidelines to tackle the anonymisation process to

[1] See https://ico.org.uk/media/1061/anonymisation-code.pdf [Accessed 13 Sept. 2018]
[2] See http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [Accessed 13 Sept. 2018].
[3] See http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2017/09/WC500235371.pdf [Accessed 13 Sept. 2018]. See also D2 for further details of Policy 0070.
[4] See https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata [Accessed 13 Sept. 2018].

minimise the risk of re-identification. It can be seen as a summary of previous content that puts the risks, techniques, and evaluation methods mentioned throughout the different deliverables into context.

It is worth emphasising that before any data is even *collected*, organisations must ensure that they have the mechanisms and safeguards in place to be data controllers. Whenever collecting data and creating databases, organisations need to be GDPR compliant, rely on third-part services that are also in line with their legal obligations (cloud services, for instance, will need to use servers based in the EU), have the necessary contracts for the handling of that data and for sharing it with data processors, and use secure mechanisms, including encryption of the information both in transit and in rest. If the raw data collected or used is sensitive,[5] if systematic and extensive profiling or automated decision-making is used, if data originates from the systematically monitoring and tracking of a publicly area or online space or data from multiples sources is combined, compared or matched, data controllers will need to conduct a Data Protection Impact Assessment (DPIA) to describe the nature, scope, context and purposes of the processing; assess necessity, proportionality and compliance measures; identify and assess risks to individuals; and design measures to mitigate those risks. Additionally, for each piece of personal data collected, data controllers must make sure that consent procedures are in place and that consent was originally given for the specific use they are planning. If reusing data, they will need to have the necessary permissions for data reuse, and again make sure that the data protection principle of purpose limitation is respected (or re-gain consent if that is not the case).

In order to help practitioners, we present the content organised in sequential *steps*. The following steps guide the analyst through the process and aim at providing support to identify potential problems and ponder potential solutions.

1. Remove obvious identifiers

Following the legal framework, no personal data, that can directly identify a person should appear as part of a public or shared dataset. Examples of these include name, family name, nickname, etc.

2. Identify potential pseudo-identifiers

The next step is to identify tuples[6] of attributes in the dataset that may act as pseudo-identifiers, i.e., that uniquely identify an individual. These tuples can be problematic in two senses. On the one hand, they may allow for the re-identification of individuals if they appear in other databases where they are linked to identifiers (e.g., the medical re-identification by Sweeney et al., or the Netflix re-identification by Narayanan and Shmatikov mentioned above). On the other hand, they also enable to link records within the database. This in turn allows for building profiles that

[5] See D1 for further details on the applicable legal framework and GDPR obligations.
[6] Tuples are data structures consisting of multiple parts.

either ease re-identification or enable inferences about individuals that would not be possible using only isolated records. Thus, pseudo-identifiers should never be made public in an unprotected way.

For the different use cases in this deliverable, there are examples of well-known pseudo-identifiers. It is important to be up to date with the literature to gain an understanding of which other attributes could become a pseudo-identifier:

- *Health*: the failures surveyed in the previous sections showed that the tuples (ZIP, Gender, DoB), (Gender, list of interventions), or genealogy trees, could be used to match individuals to other databases that gather this same information. Particular care must be given when dealing with genomic data, unique in an on itself, since it can also enable re-identification.
- *Location*: it is well known that users' movements are very unique, i.e., no two people move in the same way throughout their lives. In particular, it has been demonstrated that the tuple (home address, work address) is a pseudo-identifier that can be combined with e.g., census databases to recover names of individuals.
- *Statistics*: statistics intrinsically contain data from several people, and as such they do not contain pseudo-identifiers. Yet, as shown in the previous sections, great care must be put to make sure that enough individuals participate in the statistics so that the data does not correspond to only one person.

3. Evaluate whether other data could lead to privacy breaches

In many cases, the remaining data may still convey information about individuals. On the one hand, the remaining data can be used to rebuild the pseudo-identifier. Examples:

- *Health*: a series of medicines could be a proxy to uniquely identify a particular disease or medical procedure followed by the patient. The work by El Emam et al (2011) provides an overview of attacks that should be considered to make sure that the released data has no associated privacy risk.
- *Location*: clusters of GPS points can be used to identify points of interest, which in turn may enable the construction of pseudo-identifiers.
- *Statistics*: combination of several correlated statistics could enable to isolate data from an individual, which in turn may enable the construction of pseudo-identifiers.

On the other hand, the remaining data, even if it does not directly enable re-identification, may allow for inferences about the individuals in the database. This in turn may affect the possibility to publish the dataset in a responsible manner. Examples for the different use cases are:

- *Health*: as said above, medicines or tests can lead to infer diseases from patients.
- *Location*: when the dataset contains several people, co-locations can be used to infer social ties. Also, particular points of interest that can be found from the data can reveal religion (e.g., mosques, churches), sexual orientation (e.g., particular bars), etc.

- *Statistics*: given combinations of statistic values can not only enable isolation, but also inference about values of particular attributes for individuals.

4. Reduce privacy risks

Since there may be parts of the dataset that, following on the previous point, increase the risk of re-identification and inference, the next step is to consider possible defences that reduce this risk. These are mainly based on the techniques and methods described in Deliverable D1: randomisation, generalisation, suppression, and creation of synthetic data.[7]

Below we list possible examples for the use cases:

- *Health*: as initially proposed by Sweeney (2000), an option to avoid reidentification is the generalisation of the published values, i.e., publishing coarser quantities than the actual ones, e.g. ZIP=36XXX instead of ZIP=36491. Randomisation and suppression can also greatly help reduce the risk of re-identification. The ARX tool[8] can be used to test and visualise the impact of these defences.
- *Location*: in order to prevent inferences that eventually lead to re-identification, locations must be modified before they can be published. Popular options include the use of randomisation, which can be used following differentially private principles as described by Andres *et al.* (2013), or more advanced algorithms that account for correlations such as those proposed by Rastogui and Nath (2010). Another option is to generalise. An easy way to perform generalisation is to round the published GPS coordinates as to reduce their precision. For instance, only keeping 2 decimals gives a precision of 1.100m. Regarding suppression, depending on the application one can suppress different points, e.g., removing points at random, or removing points that are not far enough from previously reported values. The work by Boukoros *et al* (2018) provides examples of application of these mechanisms.

- Another common option is publishing aggregates to avoid having individuals' data directly published, while enabling the identification of patterns. As mentioned in D1, and above in this document, even these aggregates have the potential to reveal information about individuals. Therefore, one should think to apply similar defences to obfuscate the published aggregated values, e.g., add differentially private noise to the aggregates, round them to pre-defined values, or sample the inputs to obtain the aggregates (a form of suppression). The work by Pyrgelis *et al* (2018) provides examples of application of these mechanisms.

[7] As mentioned in D1, the latter is at a very immature stage and thus we do not provide examples.
[8] See https://arx.deidentifier.org/anonymization-tool/ [Accessed 28 Aug. 2018].

- *Statistics*: Statistics can be protected in the same ways as aggregates. Regarding differentially privacy mechanisms to protect the database, the tool DPBench by *Hay et al.* (2016) can be used to compare solutions. The code can be found online.[9]

5. Evaluate the result

As indicated in the previous documents, applying a defence in and on itself does not guarantee a null risk. It is crucial to evaluate the result of the anonymisation in order to understand both its effectiveness and the impact on utility. In particular, this evaluation is needed to tune the parameters of the anonymisation algorithm so as to obtain the best possible anonymity-utility trade-off.

As we saw above, researchers and hackers have used different methods to do this from the outside, to expose weak anonymisation. The organisation releasing a dataset should put its data through an evaluation test not only before publishing or opening data, but also regularly afterwards, as it learns about new databases (that could be used to combine with their dataset to re-identify individuals) and re-identification tools. Some actors stress the need for these evaluations and impact assessments to be done externally and independently. In this case, the database owners could rely on universities or expert consultancies to provide this independent input. Alternatively, they can internalise this task but publish the methodology and results of the test to ensure that there can be oversight over the robustness of their anonymisation techniques.

Moreover, organisations planning to release data should make sure they are GDPR compliant and have the necessary security and encryption mechanisms in place -anonymising data before its publication can become a futile exercise if the original dataset is not well protected and can be breached, as has happened in high-profile cases including the Canadian online dating service Ashley Madison[10] or the peer-to-peer ridesharing app Uber.[11]

Final remarks

All in all, our reports highlight that anonymisation is not an easy process. Data encode a lot of information, and often are strongly correlated. Thus, just deleting parts of the data is not enough to ensure that no personal data remains in the dataset.

We have provided pointers to different techniques to modify the data in order to reduce the risk of deanonymisation -both general principles and methods tailored to the health, location, and statistic use cases.

---

[9] See https://github.com/dpcomp-org/dpcomp_core [Accessed 22 Aug. 2018].
[10] See https://en.wikipedia.org/wiki/Ashley_Madison_data_breach [Accessed 22 Aug. 2018].
[11] See https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/uber-breach-exposes-the-data-of-57-million-drivers-and-users [Accessed 22 Aug. 2018].j

Yet, we stress that full anonymisation is hard to achieve. Therefore, in many cases one should accept the fact that the data is still personal and consider other possibilities to publish the data. For instance, seeking the consent of data subjects to publish the data openly and apply obfuscation to ensure that no information that these individuals would not want to make public can be inferred. Additionally, having some sort of external oversight to regularly confirm the correct use of the relevant anonymisation techniques is encouraged, in order to provide data subjects with independent guarantees in relation to how their information is managed.