Privacy and anonymization

Division for Public Administration and Development Management (DPADM), United Nations Department of Economic and Social Affairs (UNDESA).

Content

- Rationale for Anonymization
- What is personal data?
- How can individuals be identified?
- Anonymization methods and techniques
- Conclusions

Rationale for Anonymization

Governments may hold personally identifiable and commercially sensitive information. This necessarily restricts agencies' ability to share this information as open data.

WHAT OPTIONS DO YOU HAVE TO MAKE THIS INFORMATION ANONYMOUS?

- Sometimes, just removing fields from a row is sufficient.
- But sometimes this is not sufficient. It can be possible to use statistical techniques to identify individuals, especially when personal data is combined with other sources of information.

Every jurisdiction has its own requirements and every culture has its own norms about what is considered personal and private.

What is personal data?

Identifiers

- Any number or other value that is used by a computer to identify individuals can be accessed by persons with malicious intent;
- Typical identifiers include registration numbers, ID numbers, passport numbers, credit card numbers, IP addresses, and order numbers.

Very specific descriptions

- Specific descriptions make it very easy to identify individuals. As specificity increases, the number of individuals possessing a specific characteristic decreases (e.g. there are fewer French than Europeans). While specific information can be valuable, be wary of its effects in very small population segments.

An example

Census data often contains personal data that is then aggregated and anonymized using different anonymization methods. This is to ensure the aggregated data cannot be used to identify individual persons.

Compare:

“From 52,584 Migrants arriving in Germany, 22 were tested positive with HIV, half of them being from Syria.”

“From 27 Migrants arriving in 2015 in the city of Mainz, 4 came from Syria, out of them 3 were tested positive with HIV”

De-identification

De-identification is the process used to prevent a person's identity from being connected with information. Common uses of de-identification include human subject research. The aim is to protect the privacy of individuals participating in research programmes.

Common strategies for de-identifying datasets include:

- Deleting or masking personal identifiers, such as name and social security number; and

- Suppressing quasi-identifiers, such as date of birth, zip code, etc. or replacing them with information of more general nature (e.g. replacing information on the age of a group of individuals who are "22 year old" with "between 20 and 30 year old").

Re-identification

The reverse process of defeating de-identification in order to identify individuals is known as **re-identification**.

Several successful re-identifications attempts have raised doubts about the **effectiveness of de-identification** in protecting individuals' privacy. A systematic review of evidence found that published re-identification attacks were performed on data sets that were not de-identified properly (using recognized standards);

De-identification is *"somewhat useful as an added safeguard"* but not "a useful basis for policy" as *"it is not robust against near-term future re-identification methods" (*quote from United States President's Council of Advisors on Science and Technology).

Anonymization methods

Aggregate
When data are aggregated, they CANNOT be used to identify the sources of the data. FOR EXAMPLE, if we provide the mean household income for a street, we will not be able to identify the household income of any particular family.

The downside of this approach is that aggregating data too far may impair analysts' ability to interpret it.

Remove
A very simple approach to privacy protection is simply removing some fields of interest that were originally collected. For example, we could omit gender, age, location, etc.

Anonymization methods

Dithering results means to add a variation to every value within a sample, while attempting to maintain the integrity of the aggregate values. The goal is to prevent the true value for any specific value to be deduced, but to enable statistical analysis to be carried out. For example, for geographic data, one could move points of interest to a random location within a given radius.

Top and bottom coding means to replace extreme values of sensitive numerical variables with the weighted group mean for those values. This allows to mask outlying values which are potentially identifying (*e.g. if the data is about the airline industry in New Zealand, replacing results from Air New Zealand with the weighted group mean for the values in the group of data ensures that the results from this airline are not identified*).

Anonymization methods

Grouping

Multiple values can be grouped to protect individuals' privacy. Consider the following table, with the transformation appearing in the right-hand column.

132 cm 139.67 cm 143 cm 139.67 cm 144 cm 139.67 cm	144 cm 152 cm 153 cm 152 cm 159 cm 152 cm	161 cm 164 cm 167 cm 164 cm

However, there may be problems with this approach, as you will impact on the median and other percentile values.

Anonymization methods

Hash digests

Using a cryptographic hash of a string can make it impossible for someone to determine what the original string was, while allowing 3rd parties to check if strings they have are included. As they can apply the hash function to their own values, they can undertake comparisons without being able to access data that they don't already have. The transformation looks something like this:

researcher@example.org	c242dbe863aa0a38eacc72888fd41804
consumer@example.com	a99650df0d55169e0d9f1dc17194830f

Conclusions

1. The protection of personal data is critical for any OGD programme;

2. Personal data in government records needs to be removed prior to publication;

3. Information that could be used to identify individuals or small groups in population needs to be anonymized using De-identification methods;

4. If anonymization cannot safeguard personal information or information can be used to identify individuals or small population groups, data cannot be published without additional technical and legal advice;

5. The data anonymization flowchart can provide initial guidance. Technical and legal experts may need to be consulted.

Thank you!

Division for Public Administration and Development Management (DPADM) of the United Nations Department of Economic and Social Affairs (UNDESA).

OGD project and OGDCE Guidelines
publicadministration.un.org/en/ogd