Data Anonymization Guidelines

1. Introduction

The collection, storage, disclosure and use of research data among EfD research centers must comply with the EfD Data Policy and other governing rules of the respective hosting institution.

Arrangements should be prepared by the Principal Investigator (PI) or research center to carefully protect the confidentiality of participants. Before consent is obtained, researchers should inform prospective participants of:

i. Any potential risks that might mean that the confidentiality or anonymity of personal information may not be guaranteed;
ii. Which individuals and organizations, if any, will be permitted access to personal information, and under what circumstances such access will be permitted;
iii. The purpose for which personal information will be used.

Researchers must assure participants that any collected personal information that could identify them as individuals will remain strictly confidential and, depending on the research, access to the information will be restricted to the PI or to researchers directly involved in the research at all times, before, during and after the research activities. In certain types of research, where necessary and practical, personal information that could identify individual participants will remain anonymous at all times, even to the researchers themselves.

2. Data anonymization concepts

“Data anonymization” refers to the conversion or transformation of personal or confidential data into “anonymized data” by applying a range of anonymization techniques. In most cases, the process of data anonymization would be “irreversible” and the recipient of the anonymized dataset would not be able to recreate the original data.

Wherever possible data should be collected, stored or handled in anonymous form. Where linkage between datasets is required (e.g. in longitudinal studies) record numbers should be used as far as possible, with special measures used to protect the key that would link a number to personal identifiers.

3. Basic data anonymization techniques

The following anonymization techniques can be used depending of the nature of the data to be anonymized and specific purpose of the anonymization. Researchers are advised to study the details of any of these anonymization techniques before using them.

- Attribute suppression
- Record suppression
- Character masking
- Pseudonymization (coding)
- Generalization
- Swapping/shuffling/permutation
- Data perturbation
- Synthetic data
- Data aggregation

Data Anonymization Guidelines

It must be recognized that there is no “one size fits all” solution for EfD centers or any organization. Each center should therefore utilize anonymization approaches that are appropriate for their circumstances. Some factors that organizations can take into account when deciding which anonymization techniques to use include:

- The nature and type of personal data that the organization intends to anonymize, as different anonymization techniques are suitable for different types of data and circumstances;
- Risk management by the organization to impose controls to protect the anonymized data, in addition to the anonymization techniques;
- The utility required from the anonymized data.

4. Anonymization procedures and tools

There are many tools or software that can be used to anonymize data such as ARX, MU-ARGUS, Anonymizer, etc. ARX is the most common and comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of privacy and risk models, methods for transforming data and methods for analyzing the usefulness of the resulting data. EfD researchers can use any anonymization tools convenient for them.

5. Disclosure risks

There are three different types of disclosure risks: identity, attribute and inference disclosure risks. Identity disclosure involves determining, with high level of confidence, the identity of an individual described by a specific record. On the other hand, attribute disclosure is determining that an attribute described in the dataset belongs to a specific individual, even if the individual’s record cannot be distinguished, while inference disclosure means making an inference about an individual even if he/she is not in the dataset, by statistical properties of the dataset. As supported by most traditional anonymization techniques, researchers should at least aim to protect against identity disclosure risk.

For any questions related to anonymization of research data or other data management related issues, don’t hesitate to contact data@efd.gu.se, marleen.poot@ub.gu.se or samuel_abera@yahoo.com