Office for National Statistics

ons.gov.uk

De-identification policy

De-identification policy

Scope

This policy provides an outline of the principles followed when de-identifying data for the production of statistics and statistical research. The policy document itself does not aim to identify best practices in de-identification or describe the relevant processes.

This policy applies to all Office for National Statistics (ONS) staff producing, using, or disseminating de-identified data as well as to internal and external parties accessing de-identified data on ONS systems. It applies to all data that identify directly or indirectly a person or business, whether the ONS is the controller or the processor.

Background

The Office for National Statistics (ONS) relies on an increasing number of data sources in multiple domains to facilitate the production of statistics and statistical research. Making high quality data available for research is an integral part of the statistical system, empowering researchers to deliver innovative research projects, validate and constructively challenge research and statistical outputs.

When working with de-identified data is feasible, a proportionate level of de-identification ensures that ONS staff and accredited researchers can access the data they require without directly or indirectly disclosing the identity of individuals (re-identification). Mitigating the re-identification risk is achieved by applying a holistic set of controls to minimise this risk. The requirement for statistical disclosure controls (SDC) further assists by ensuring published outputs are no longer personal information.

Failing to apply such controls would have significant consequences for our reputation, reducing the public's trust in official statistics and statistical research. It will also limit our capability in using data securely, legally, and ethically to deliver and facilitate research, which presents a clear benefit to the public. Finally, it will also damage our relationships with data suppliers significantly restricting collaboration in the production of statistics and statistical research and limiting our capabilities for data acquisition and sharing.

In addition to reputation risk, failure to put in place robust safeguards to reasonably prevent the re-identification of data subjects during analysis and disclosure presents a legal risk, as we fail to meet our legal requirements to ensure the safety and privacy of data subjects (individuals and businesses).

Consequently, our capability to introduce, test and review controls is essential in all statistical and research environments to maintain our accreditation status. Withdrawal of the accreditation status under the Digital Economy Act (DEA) for the ONS would not only lead to a reputational and public trust risk, but harm our relationship with data suppliers. The latter would curtail our capability to acquire, use and curate data for the production of statistics and statistical research. This could lead to operational ramifications, especially

for domains depending on de-identified data, such as the Integrated Data Service (IDS), the Secure Research Service (SRS), and the Data Access Platform (DAP).

As part of the ONS's commitment to the highest ethical standards, we are taking a leading role in safeguarding confidentiality by ensuring the efficacy and proportionality of controls applied to de-identify data. We also commit to transparency in how we make de-identified data available for statistical research in the public interest.

Policy statement

Data de-identification is conducted by the Office for National Statistics (ONS) to ensure data minimisation and to protect confidentiality when producing statistics, undertaking research, or enabling access to de-identified data to accredited researchers and projects.

The ONS is responsible for ensuring that reasonable de-identification is done effectively, securely, legally, and ethically, including privacy considerations and complies with well-established quality standards. The risk of re-identification and equivalent de-identification practices should be considered at each stage of the data journey for each dataset, on a case-by-case basis.

Where possible, commonly agreed methods and quality metrics shall be adopted to promote a harmonised approach to de-identification across government. As new methods emerge, the ONS should assess if these are suitable and develop guidance on their implementation.

Policy detail

De-identified data are a critical part of the Office for National Statistics' (ONS's) data estate, allowing us to produce analysis with minimal risk of disclosure during analysis and dissemination. De-identified data remain easier to acquire, access within a secure environment, and share. Nonetheless, de-identified data should not under any circumstances be regarded as safe or risk-free. Depending on multiple factors, any de-identified dataset entails a residual risk of re-identification. Examples include, the de-identification methods, the variable selection, the methods of access and sharing, the security arrangements and the potential to cross-reference using multiple datasets.

As a result, the process of de-identifying data consists of:

- evaluating the re-identification risk at all stages of the data life cycle
- agree on the risk tolerance for re-identification (how much risk are we willing to accept), taking into account any legal requirements for data use and access
- process the data and/or introduce controls (security, privacy enhancing, statistical) to limit that risk
- determine if the residual re-identification risk is acceptable
- review the residual risk and the risk tolerance

Particular emphasis should be given when the access and use of a dataset requires de-identification mandated by the data controller. In such cases, it is imperative that we meet all requirements set by the data supplier, as methodology adopted or curation of original and processed data.

The data sensitivity model can offer an indication of the level or re-identification risk by exploring the characteristics of the data at a specific point of access. The data sensitivity score should not be used as a sole indicator of the risk of de-identification, in most analytical environments multiple datasets are made available at the same time and data are matched or linked, which significantly compounds the risk of re-identification. The scope of research projects (for example, particular geographies and unique characteristics) and external factors (for example, publicly available data and information, or inside knowledge of a sector or research area) contribute to escalated risk. Finally, the level of the residual re-identification risk and the respective tolerances vary at different stages of the data life cycle.

The range of factors affecting the mitigation of the re-identification risk makes it difficult to guarantee that a data subject is not identifiable at all stages of the data life cycle. For example, we cannot completely mitigate the risk of a data user discovering the identify of a data subject by using external information at their disposal or personal knowledge of the data subject during analysis. We can, however, protect this information from being disclosed by applying a framework of controls, as the 5-Safes framework. Such frameworks are based on a combination of technical and security controls in addition to strict conditions of accreditation.

The main considerations for de-identifying data are presented in Annex A: De-identification through the data journey.

Roles and responsibilities

ONS staff carrying out data de-identification, predominantly the Data Engineering team

Responsible for:

- complying with the data de-identification policy
- following best practice in de-identifying data
- addressing quality or technical issues regarding the de-identification and processing of data
- reporting incidents and breaches to the ONS Data Security and Data Protection team respectively

Researchers accessing de-identified data

Responsible for:

- complying with all policies in the data hosting environment
- applying disclosure controls consistently to all statistical and research outputs
- reporting any concerns in terms of quality and re-identification to the data engineering (for ONS data) or the support team (for other domains for example, the

Integrated Data Service (IDS), the Secure Research Service (SRS), and the Data Access Platform (DAP))
- reporting incidents and breaches immediately to the support team

Domain support staff, including IDS, SRS, and DAP support staff

Responsible for:

- storing and curating de-identified data, along with the relevant metadata, securely and consistently
- ensuring that disclosure controls are applied consistently to all statistical and research outputs
- reporting any concerns in terms of quality and re-identification to the data engineering team (for ONS data) and/or the data provider
- investigating reported incidents and breaches to the ONS Data Security and Data Protection team respectively

Information Asset Owner (IAO) or Data Steward

Responsible for:

- ensuring that data asset is registered on the Information and Data Register or Data Catalogue, the record is accurate, and the data sensitivity and security classification is correctly assessed
- making decisions around data access and use of the dataset as and when required

ONS Data Security

Responsible for:

- reviewing the security arrangements across the different data hosting platforms and issuing relevant actions
- providing an assurance statement to the Data Governance Committee and the National Statistics Executive group regarding the security of de-identified data and escalating any concerns

ONS Data Ethics

Responsible for:

- reviewing ethics applications and ethics self-assessments to assess the ethical risks when using de-identified data for the production of statistics and statistical research
- producing guidance and offering advice on managing ethical risks
- providing assurance to the National Statistician's Data Ethics Advisory Committee that ethical risks are thoroughly considered upon approval

National Statistician's Data Ethics Advisory Committee

Responsible for:

- considering ethics applications to assess the ethical risks when using de-identified data for the production of statistics and statistical research
- providing assurance to the National Statistician that ethical risks are properly considered and managed

Data Governance Committee

Responsible for:

- ensuring security risks stemming from accessing, using, and sharing de-identified data are properly managed and escalated as and when required
- assessing how practices around the curation of de-identified data affect the overall risk exposure of the organisation
- providing approval, advice, or steer in policy decisions and data management considerations