Glossary of Terms

Anonymisation: Anonymisation is a complex process that transforms personal identifiable data into non identifiable (anonymous) data. This requires that identifiers be removed, obscured, aggregated and/or altered in some way. There are two types of identifiers: formal/direct identifiers and complex identifiers. Formal identifiers include such as a data subject's name, address and unique reference numbers e.g. their social security number or National Health Service number. Complex identifiers can in principle include any piece of information (or combination of pieces of information). For example, consider the following combination of information a "sixteen year old widow", whilst *age* and *marital status* are not immediately obvious identifiers, our implicit demographic knowledge tells us that this is a rare combination. This means that such an individual could potentially be re-identified by for example someone spontaneously recognising that this record corresponded to someone they knew.

Data divergence: This represents the differences between two datasets (data-data divergence) or between a single dataset and reality (data-world divergence). Sources of data divergence include: data ageing, response errors, coding or data entry errors, differences in coding and the effect of disclosure control.

Data intruder: A data user who attempts to disclose information about a data subject through identification and/ or attribution (see statistical disclosure). Intruders may be motivated by a wish to discredit or otherwise harm the organisation disseminating the data, to gain notoriety or publicity, or to gain profitable knowledge about particular data subjects.

Data protection: This refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into data subjects' privacy caused by the collection, storage and dissemination of data.

Data subject: A respondent in a dataset.

Data utility: A term describing the value of a given data release as an analytical resource - the key issue being does the data represent whatever it is that it is supposed to represent? Disclosure control methods can have an adverse effect on data utility. Ideally, the goal of any disclosure control regime should be to maximise data utility whilst minimising disclosure risk. In practice disclosure control decisions are a trade-off between utility and disclosure risk.

Disclosure control methods: These can be defined as a set of methods for reducing the risk of disclosure such methods are usually based on restricting the amount of, or modifying the, data released.

Disclosive data: Data are considered to be disclosive when they allow data subjects to be identified, (either directly or indirectly) and/or when they allow information about data subjects to be revealed.

Disclosure risk: This is expressed as the probability that an intruder identifies and or reveals new information about at least one data subject in the disseminated data. Because anonymisation is difficult and has to be balanced against data utility, the risk that a disclosure will happen will never be zero. In other words there will be a risk of disclosure present in all useful anonymised data.

Formal identifier: Sometimes referred to as direct identifier, include such as a data subject's name, address and unique reference numbers e.g. their social security number or National Health Service number.

Informed consent: Basic ethical tenet of scientific research on human populations. Informed consent refers to a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including awareness of any risks involved, of uses and users of the data, and of alternatives to providing the data.

Key variable: A variable in common between two datasets, which may therefore be used for linking records between them.

Licensing agreement: A permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information.

Microdata: A microdata set consists of a set of records containing information on individual data subjects.

Personal data: Any information relating to an identified or identifiable data subject. An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous.

Population unique: A record within a dataset which is unique within the population on a given set of key variables.

Privacy: Privacy is a concept that applies to data subjects while confidentiality applies to data. There is a definite relationship between confidentiality and privacy. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used. Informational privacy encompasses an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

Record linkage process: A process attempting to classify pairs of matches between different datasets.

Remote access: On-line access to protected microdata.

Restricted access: A data protection measure that limits who has access to a particular dataset. Approved users can either have: (i) access to a whole range of raw (protected) data and process it themselves or (ii) access to outputs e.g. tables from the data.

R-U map: A graphical representation of the trade-off between disclosure risk and data utility.

Safe data: Data that has been protected by suitable Statistical Disclosure Control methods.

Safe setting: An environment such as a data lab whereby access to a disclosive dataset can be controlled.

Sample unique: A record within a data set which is unique within that dataset on a given set of key variables.

Sampling: This refers to releasing only a proportion of the original data records on a microdata file. In the context of disclosure control, a data intruder could not be certain that any particular person was in the file.

Sampling fraction: The proportion of the population contained within a dataset. With simple random sampling, the sample fraction represents the proportion of population units that are selected in the sample. With more complex sampling methods, this is usually the ratio of the number of units in the sample to the number of units in the population from which the sample is selected.

Scenario analysis: A framework for analysing plausible data intrusions attempts. This framework identifies (some) of the likely factors, conditions and mechanism for disclosure.

Sensitive variables: Variables contained in a data record that belong to the private domain of data subjects who would not like them to be disclosed. There is no exact definition given for what is a 'sensitive variable'. Some data are clearly sensitive such as the possession of a criminal record, one's medical condition or credit record, but there are other cases where the distinction depends on the circumstances, e.g. one's religion might be considered as a sensitive variable in some countries and not so in others.

Statistical Disclosure Control (SDC): SDC is an umbrella term for the integrated processes of disclosure risk assessment, disclosure risk management and data utility.

Statistical disclosure: A statistical disclosure is a form of data confidentiality breach that occurs when, through statistical matching, an individual data subject is identified within an anonymised dataset and /or confidential information about them is revealed. A statistical disclosure may come about through: (i) the processes of identification and attribution (i.e. the revealing of new information) or (ii) the process of attribution alone.

Synthetic data: Data that have been generated from one or more population models.

Tabular data: Aggregate information on entities presented in tables.

Target dataset: An anonymised dataset which is used by an intruder to attempt to identify particular data subjects.