CLS Data Classification Policy

Document Information	
Document Name	CLS Data Classification Policy
Author(s)	Aida Sanchez
Version	3
Issue Date	September 2020
Approved By	CLS Data Access Committee
Next review	Yearly

Document History		
Version	**Date**	**Summary of change**
0.1	30/04/2014	Initial issue
1.2	01/05/2014	Amendments post review (JJ)
1.3	18/06/2014	Final review (JJ) & format amendments (GG)
1.4	01/10/2014	Review (GG & HW)
1.5	03/10/2014	Updated with review feedback
1.6	27/10/2014	Updated, minor corrections
1.7	14/11/2014	Updated inc. formatting
1.8	27/02/2015	Version presented to IGSG
2.0	27/02/2105	Approved by IGSG (inc. rebranding)
2.1	June 2018	Table of data classification updated (AS)
3	September 2020	Reformatting; update of the data classification table; addition of data classification principles, data disclosure methods and description intruder profiles

This document includes data that is **PUBLIC** and can be disclosed outside UCL CLS and used, or disclosed in whole or in part for any purpose other than to evaluate and implement procedures defined within this document.

Contents

1. Introduction

The CLS Data Classification Policy defines the principles through which the CLS data is categorised, and describes the information classification categories which will be applied to all CLS data and information.

2. Principles of data categorisation

CLS data have been collected or linked to external data sources to support research by a wide range of researchers in universities and other settings. The data collection, linkage, dissemination and sharing of data is based on the consent given by the participants and is conditional on the promise that we will protect their confidentiality and other rights in relation to the data. Attempts to re-identify individuals in the cohort is always forbidden.

Breaking this promise would not only constitute an ethical violation of consent, but also threatens the trust that cohort members place in the research team who collect their data, and may affect their willingness to participate in further data collection.

CLS have evaluated and categorised the data in terms of three underlying principles:

- Disclosivity
- Sensitivity
- Potential consequences of misuse of data

The appropriate degree of security and access management will be applied depending on what category has been assigned.

2.1 Disclosivity

Data are considered disclosive if there are concerns over the re-identification of individuals, households or organisations with which they are associated, should data users attempt to do so.

CLS data is categorised reflecting an assessment of the likelihood and potential impact of disclosure of individual.

2.2 Sensitivity

Sensitive data are those that require more protection because of their content, and participants might reasonably expect that additional steps are taken to ensure disclosure risk was reduced.

Data are considered sensitive if there are concerns over the consequences of re-identification, i.e., the potential damage in the case of self-identification or identification by other family members and also because participants may expect such data to be subject to greater protection. For instance a participant may be concerned about their drinking behaviour being discoverable but not their ethnic origin.

Information on detailed mental or physical health, illegal behaviour, childhood abuse, drug/alcohol use is considered particularly sensitive. Other sensitive data include racial or ethnic origin, religion, genetics or sexual orientation. Such data are considered "special category data" according to the Data Protection Act 2018 categorisation (https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/).

Data that risk the disclosure of sensitive information requires a high degree of security and management.

2.3 Potential consequences of the misuse of data

The consequences of misuse depend on a number of factors, including whether it is accidental or malicious, the scale of data disclosure (i.e. how many participants it affects), whether it creates a possibility of harm or distress for participants or whether it affects the public reputation of the study. Consequences may include negative publicity and legal action.

In terms of categorisation, the impact of consequences are closely aligned with the other principles. For instance, sensitive data should be classified at a higher impact level because the consequences of misuse are more severe.

Risk of disclosure and misuse has two components: a) the risk that a user of the data attempts to breach the confidentiality of participants or misuses the data in any other way and b) the risk that they are able to do so given the data they available. The UK Anonymisation Network recommend considering who might try to re-identify the individuals in a dataset (here-on referred to as intruders). These would be users who they are willing to sign up to the licence but not comply with it. Appendix 1 outlines the template for an intruder profile.

3. Data classification scheme

Data held at CLS is classified under the following schema, which determines the level of access:

- **Public**: Open Public access
- **Restricted**: Controlled public access. There are four levels of access restrictions, which divide the data into four sub- groups known as ‘tiers’ and based on the level of sensitivity and potential disclosure:
    - **Tier 1**: low level of disclosure and sensitivity
    - **Tier 2a**: medium level of disclosure and/or sensitivity
    - **Tier 2b**: high level of disclosure and/or sensitivity
    - **Tier 3**: very level of disclosure and/or sensitivity
- **Confidential**
- **Private**

The data classification assigned to a particular set of data determines the levels of public access.

The details of these data categories are summarised in the table below, and provides examples of the different kinds of data and information which would be covered within the agreed categories.

Placing data in higher categories provides greater protection for participants but increases the real or perceived barriers to use of the data by researchers. Real barriers include limitations on access outside the UK for higher impact level data. There is therefore a balance to be drawn between maximising the use of the data and minimising risks to the rights of participants.

This table illustrates the data classification based on their level of disclosure risk and how the data are publicly accessible.

Classification	Description	Disclosure risk	Sensitivity	Access	Approved users
PUBLIC	Publicly available datasets. Example: Edubase, list of schools (http://www.education.gov.uk/edubase/search.xhtml)	None	None	Public websites	
RESTRICTED - TIER 1	Pseudo-anonymised and de-identified survey data with low level of disclosure. Example: the Age 46 follow-up of BCS70 https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8547	Low	Low	UKDS End User Licence	Registered UKDS users
RESTRICTED - TIER 2a	Data with a medium level of potential disclosure risk (e.g intermediate geographical indicators such as counties) or sensitivity (e.g mortality data, detailed physical or mental health information, genetics). Example: the NCDS counties https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=5744	Medium	Low/Medium	UKDS Special Licence or METADAC (genetics)	EEA research projects approved by UKDS/CLS or METADAC (genetics)
RESTRICTED - TIER 2b	Data with a high level of potential disclosure risk (e.g linked education data, exact dates, detailed ethnicity, detailed geographical indicators such as OA, LSOA,	High	Medium/High	UKDS Secure Access	UK based research projects

	MSOA, Local authority) or highly sensitivity such as linked health data or education data. Example: NPD data linked to Next Steps https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7104				approved by UKDS/CLS
RESTRICTED - TIER 3	Data with a very high level of potential disclosure. Any information which would allow identification of less than 5% of a population of the data item Example: open text responses, postcodes, school IDs	Very High	Medium/High	UCL Data Safe Haven	CLS data managers and UK researchers approved by CLS DAC
CONFIDENTIAL	Individually identifying information only accessible for operational purposes: contact with cohort members, data collection, external data linkage. Example: names, address, email, NHS Number, National Insurance Number (NINO), Unique Pupil Number (UPN), Unique Learner Number (ULN), etc.	Full identification	Low	UCL Data Safe Haven and highly secure external servers as needed	CLS cohort maintenance team and external parties as required
PRIVATE	CLS internal documentation for which there is no benefit or requirement to make it publicly available	n/a	n/a	CLS shared drive on UCL server	CLS staff

4. Methods of disclosure control

CLS follow the ONS and UKDS guidelines of disclosure control:

https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata

https://www.closer.ac.uk/wp-content/uploads/Methods-of-disclosure-control-the-UKDS-approach-to-review-Louise-Corti.pdf

We assess disclosure risk based of potentially disclosive variables taking into account that:

- there isn't an exact formula to help us judge 'objective' risk
- we cannot give a one-size-fits-all rule book
- we follow the recommended best practice for surveys

4.1 Identifiers used for pseudo-anonymisation

CLS data is held and distributed in a pseudo-anonymised manner, which consists of assigning different identifiers to the data. This ensures that if data is released, linkage to other data is only possible to other data on the same identifier. For instance, participants' contact data used for data collection fieldwork or for matching to external administrative data should not be linkable to research data available from the UK Data Service.

CLS Internal identifier

These are identifiers used on databases and files in both the CLS Research Data Management team and the CLS Cohort Maintenance team. This is to allow appropriate sharing on data across these two functional areas.

These identifiers or data linked to them or lookups between this and other identifiers should *never* be released from CLS.

Data collection identifier

These are identifiers for use during fieldwork or other data collections where data is passed to a third party for contact with cohort members.

This also includes research groups who want to recode string data for other purposes e.g. re-coding of occupation or diseases for coding frames not currently available. The research collaborator would need to return the coded data for it to be relinked to a research identifier before it could be merged in with existing data.

Research identifiers

These are identifiers used by researchers and deposited at the national repositories such as the UK Data Service. Examples of these are

- NCDS : NCDSID
- BCS70 : BCSID
- Next Steps: NSID
- MCS : MCSID

In some studies, for instance NCDS, where there is linkage between data collected during a survey and derived data from genetic analysis, a bespoke research identifier is issued.

4.2 Methods to reduce data disclosivity

Assessment of disclosure risk of potentially disclosive and sensitive variables is performed following recommended best practices for surveys and the ONS and UKDS guidelines of disclosure control. Once the potential disclosivity has been ascertained, CLS data are checked and suitably de-identified prior to disseminating for research purposes. The depth of the de-identification applied will depend of the chosen mechanism for data access, i.e, the data classification assigned.

Assessment of the data disclosivity

Examples of detailed and/or sensitive information that could potentially lead to the identification of individual or households include:

- Exact dates: birth date, data collection date
- Detailed employment (full SOC/SIC)
- Religious affiliation, ethnicity, language spoken at home, country of origin
- Outliers (e.g. height, number of bedrooms, number of children)
- Very detailed health information (e.g. full ICD or BNF codes)
- Unusual health condition (e.g. rare disease, total blindness)
- Small geographic area such as postcodes, OA, LSOA, etc
- Local neighbourhood specific characteristics (e.g. detailed IMD)

- Open ended answers in qualitative research
- Linked information: school identifiers, health care providers, etc

CLS have develop a number of programming scripts and protocols to find sensitive and disclosive variables and to tabulate ‘risky’ variables with small cell counts. For instance, the threshold chosen is to have no cell counts less than 10 for data released under the Restricted Tier 1 (UKDS End User Licence). Where possible, variables are also be checked against the population from which the sample was taken (such as height distributions).

Data disclosure methods applied to enable data sharing

Once the potentially disclosive variables have been identified, CLS applies a number of modifications to the data in order to create a dataset that can be publicly shared under the chosen data sharing method.

Some of these de-identification methods applied by CLS are:

- **Banding** – reduced granularity of information whilst retaining some information of the distribution.
- **Top/bottom coding** – where a continuous distribution has a long right or left tail (or both), those outliers are assigned a maximum value (top-code) or minimum value (bottom-code) so that they are grouped together at the top and bottom of the scales.
- **Reducing precision** –this could be in the form of truncation, such as only providing the first half of a postcode or the first three digits of a SOC code..
- **Pseudonymisation** – there may be indirect identifiers which also act as “clusters”, such as GP practice, or school. These can be given a code which retains the clustering but removes the identifying information.
- **Variable removal** – where a variable is considered too disclosive or sensitive for the intended licence.

Appendix 1. Intruder profiles

The UK Anonymisation Network (UKAN, http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf) recommend considering who might try to re-identify the individuals in a dataset (here-on referred to as intruders). For the purpose of the exercise it is assumed that they are willing to sign up to the end user licence but not comply with it (as previously mentioned, attempted to re-identify individuals in the cohort are forbidden). Table 1 outlines the template for an intruder profile.

Searching for external data, such as the electoral register, social media (for example, educational qualifications may appear on LinkedIn) and look-ups for codes. If the data can be linked, then the disclosure check should include that data available. The attack profiles help build a picture of what may be available to an attacker and what could be linked (see above).

Three intruder profiles have been identified:

- **Activist** – a group or individual who wishes to discredit data sharing, the NCDS, the centre for longitudinal studies, University College London or the Institute for Education
- **Marketer**- Companies who wish to use the large dataset for some marketing purpose. If one could re-identify a proportion of the cohort members who were found to have a profile of particular biometric profile that would be useful for marketing, particularly with income information and any geographical information.
- **Nosy Neighbour** - i.e. someone who knows a cohort member & that they are a cohort member. This could be someone who is aggrieved with the cohort member and wishes to bring them to disrepute

Table 1: Template for putting together an intruder profile, from Elliot & Dale (1999)[i]

INPUTS	
Motivation:	What are the intruders trying to achieve?
Means:	What resources (including other data) and skills do they have?
Opportunity:	How do they access the data?
Target Variables:	For a disclosure to be meaningful something has to be learned; this is related to the notion of sensitivity.
Goals achievable by other means?	Is there a better way for the intruders to get what they want than attacking your dataset?
Effect of Data Divergence:	All data contain errors/mismatches against reality. How will that affect the attack?
INTERMEDIATE OUTPUTS (to be used in the risk analysis)	
Attack Type:	What is the technical aspect of statistical/computational method used to attack the data?
Key Variables:	What information from other data resources is going to be brought to bear in the attack?
FINAL OUTPUTS (the results of the risk analysis)	
Likelihood of Attempt:	Given the inputs, how likely is such an attack?
Likelihood of Success:	If there is such an attack, how likely is it to succeed?
Consequences of Attempt:	What happens next if they are successful (or not)?
Effect of Variations in the Data Situation:	By changing the data situation can you affect the above?

[i] Elliot M.J. & Dale A. (1999) Scenarios of Attack: The Data Intruder's Perspective on Statistical disclosure risk. Netherlands Official Statistics, Special Edition, Spring