National
Coordination Point
Research Data
Management

Risk management for research data about people

A GENERAL MATRIX TO BE USED BY DATA STEWARDS AND RESEARCH SUPPORTERS FOR THE ASSESSMENT OF PRIVACY RISKS WITH RESEARCH DATA AND DETERMINATION OF APPROPRIATE METHODS FOR RISK MANAGEMENT

'How likely is it that a person can be re-identified from research data?'

'What are appropriate measures to protect the data and the people behind the data?'

What?

This matrix is generic. It is a tool for data stewards or other research supporters to assist researchers in taking appropriate measures for the safe use and protection of data about people in scientific research. It is a template that you can adjust to the context of your own institution, faculty and / or department by taking into consideration your setting's own policies, guidelines, infrastructure and technical solutions. In this way you can more effectively determine the appropriate technical and organizational measures to protect the data based on the context of the research and the risks associated with the data.

Why?

Data about people used in scientific research are rarely anonymous. It is important that researchers are aware of this at an early stage in their data management planning because, if data are not anonymous, the General Data Protection Regulation (GDPR) applies. This means, amongst other things, that the correct technical and organizational measures must be taken to protect these data.

How?

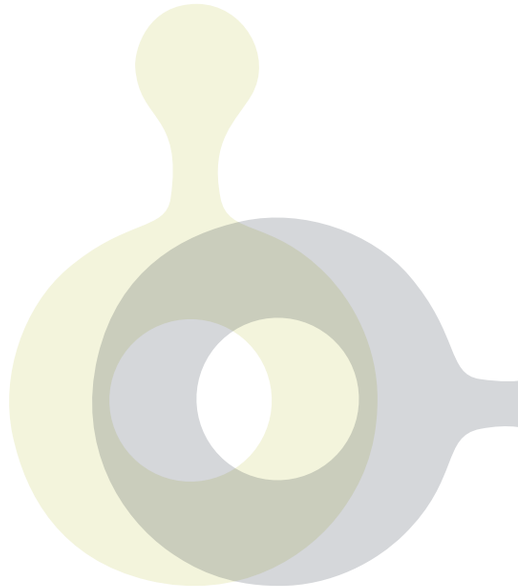
The matrix is based on the *Five Safes Framework*[1]. This framework consists of five perspectives (projects, people, data, settings, and output) that should be considered when determining appropriate data protection measures. In general, data protection measures should address all five of these perspectives. If certain measures in one perspective are not feasible, "stricter" measures should be applied in another perspective of the *Five Saves Framework*.

One data protection measure that is particularly important with research data is pseudonymization. The level of de-identification can vary with pseudonymized data, but *always remember*: even if data are pseudonymized, they are still **not** anonymous. The matrix below provides guidance on data protection measures that can be used for the various de-identification levels of pseudonymized data while taking into consideration each perspective of the *Five Safes Framework*.

In summary, this matrix helps you in your role of a research supporter to provide discipline-specific advice on data protection measures that is in line with the GDPR requirements and also consistent with other research institutions in the Netherlands.

The matrix provides the following information:

I. 5 risk levels for how likely or easily an individual could be identified from the data
II. A generic example for each of these levels
III. A field that should be completed with discipline specific examples
IV. 5 perspectives from the *Five Safes Framework* to consider for each risk level.



And finally

If you are in doubt or if you have questions about anonymization, pseudonymization and data protection measures, always talk to the privacy officer in your own institution.

LCRDM Task Group Anonymization
(Version: December 2019)

1] Meer informatie over
The Five Saves Matrix (Engels):
– https://en.wikipedia.org/wiki/Five_safes;
– https://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf
– http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/
– http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/keep-data-safe.aspx

Following guidance is recommended as a general framework for each institutions' data protection measures.
Add additional detail to each section as required.

SECTION I: What are the risk levels for re-identification of research data about people?

Identification risk level	**PS0** Not pseudonymized	**PS1** Pseudonymized at level 1	**PS2** Pseudonymized at level 2	**PS3** Pseudonymized at level 3	**"ANON"** Anonymized
How high is the risk of re-identification?					
Definition for each risk level	*Directly identifying∗ personal data*	*Direct identifiers∗ are not present but:* *1. participants are identified via pseudonym/ ID number that links to a linking table/key file∗∗ containing the directly identifying information* ***AND*** *2. the pseudonym or ID number is meaningful or not random, e.g. DOB + postal code* ***AND/OR*** *3. data collected can easily be used to re-identify an individual*	*Direct identifiers∗ are not present but:* *1. participants are identified via a meaningless pseudonym/random ID number that links to a linking table/key file∗∗ containing the directly identifying information* ***AND/OR*** *2. a unique profile for an individual could be generated from the collected data* ***AND/OR*** *3. with some reasonable time and effort, it is possible to re-identify an individual based on characteristics in the data*	*Direct identifiers∗ are not present but* *1. participants are identified via a meaningless pseudonym/random ID number that links to a linking table/key file∗∗ containing the directly identifying information* ***AND*** *2. it is not possible to generate a unique profile for at least one individual from the collected data* ***AND*** *3. it would not be possible to re-identify an individual based on characteristics in the data*	*Data collected anonymously:* *1. No direct or indirect identifiers∗ present.* ***AND*** *2. There is no linking table/key file possible∗∗ (i.e. there is no way to couple the anonymous data to another dataset)* ***AND*** *3. Insufficient data is collected to create a profile unique to at least one individual* ***AND*** *It is not possible to re-identify participants based on characteristics in the data*

* Directly identifying/direct identifiers: Data can be directly and easily attributed to an individual through characteristics/variables that are unique to that individual such as name, address, e-mail address, BSN etc. Note that directly identifying variables may depend on the context or the individual in question (e.g. Jan Janssen versus Mark Rutte), so you may need to consider the context when deciding if a variable is directly identifying.
→Indirectly identifying/indirect identifiers: Data that must be combined with other information to identify an individual, such as a random ID code that links to directly identifying information or through a combination of variables that singles out a unique individual (e.g. a man in a breast cancer registry can be identified through combination of gender and breast cancer status)

** Linking table/key file: a dataset containing directly identifying information that is linked to research data via a random ID code.

There may be identifiers in the dataset with which people can be identified via a different file. In this way you also create a key file, unintentionally. For example, if the dataset contains technical keys (record IDs) from a source file or if the dataset contains numbers/IDs for lab samples/measurements or document IDs or filenames.

NB: This level applies, even if no random ID numbers linking to a linking table/key file are used, but it is still possible to generate unique profiles of individuals based on the collected data and/or it would be possible to re-identify an individual based the characteristics collected in the data

NB: At this level of pseudonymization, the data are very close to being anonymous, but ultimately the GDPR still applies. There may be situations where the data could be handled in a similar manner to "ANON"-level data as long as the linking tables/key files are kept extremely secure, and ONLY after discussion and agreement from privacy and security experts at your institution.

Following guidance is recommended as a general framework for each institutions' data protection measures.
Add additional detail to each section as required.

SECTION II: Generic example of research data at each identification risk level

PS0 Not pseudonymized	**PS1** Pseudonymized at level 1	**PS2** Pseudonymized at level 2	**PS3** Pseudonymized at level 3	**"ANON"** Anonymized
NAME: Rutger Hauer PATIENT NUMBER: 90210 E-MAIL: blade.runner@batman.nl POSTAL CODE: 8911 AA CITY: Leeuwarden DATE OF BIRTH: 27-4-1967 INCOME: 7,861 JOB: Judge CAR: DeLorean LICENSE PLATE: SN-09-HN	PATIENT NUMBER: 90210 POSTAL CODE: 8911 CITY: Leeuwarden DATE OF BIRTH: 27-4-1967 INCOME: 7,861 JOB: Judge CAR: DeLorean LICENSE PLATE: SN-09-HN	STUDY SUBJECT: 47110009 REGION: Friesland YEAR OF BIRTH: 1967 INCOME: 7,500-10,000 JOB: Legal CAR: DeLorean	STUDY SUBJECT: 47110009 COUNTRY: Netherlands AGE: 51-60 INCOME: 5,000-15,000 JOB: Legal CAR: Sports car	COUNTRY: Netherlands AGE: 51-60 INCOME: 5,000-15,000 JOB: Legal CAR: Sports car

SECTION III:
Discipline specific research data examples to be filled in by users of the matrix

Users should add additional discipline specific examples	*Users should add additional discipline specific examples*	*Users should add additional discipline specific examples*	*Users should add additional discipline specific examples*	*Users should add additional discipline specific examples*

SECTION IV:
Five perspectives (projects, people, data, settings, output) to consider when determining data protection measures

	PS0 Not pseudonymized	**PS1** Pseudonymized at level 1	**PS2** Pseudonymized at level 2	**PS3** Pseudonymized at level 3	**"ANON"** Anonymized
Safe projects *How to ensure use of the data is appropriate, legal and ethical?*	Researcher should: – Complete a DPIA, DMP and ethics application ***prior to data collection*** – Check whether informed consent is required and whether the consent process has been followed. – Ensure legal agreements between involved parties that are required by GDPR are in place ***prior to data collection*** – Check if other legal agreements are required for business confidentiality or intellectual property purposes	Researcher should: – Complete a DPIA, DMP and ethics application ***prior to data collection*** – Check whether informed consent is required and whether the consent process has been followed. – Ensure legal agreements between involved parties that are required by GDPR are in place ***prior to data collection*** – Check if other legal agreements are required for business confidentiality or intellectual property purposes	Researcher should: – Complete a DPIA, DMP and ethics application ***prior to data collection*** – Check whether informed consent is required and whether the consent process has been followed. – Ensure legal agreements between involved parties that are required by GDPR are in place ***prior to data collection*** – Check if other legal agreements are required for business confidentiality or intellectual property purposes	esearcher should: – Complete a pre-DPIA (to see whether or not a DPIA is necessary), as well as DMP and ethics application ***prior to data collection*** – Check whether informed consent is required and whether the consent process has been followed. – Ensure legal agreements between involved parties that are required by GDPR are in place ***prior to data collection*** – Check if other legal agreements are required for business confidentiality or intellectual property purposes	Research should: – Complete a DMP and ethics application, where applicable, ***prior to data collection*** – Check with experts to confirm that data are in fact anonymous; choose experts appropriate to your discipline that can appropriately assess the type of data in question. – Check if other legal agreements are required for business confidentiality or intellectual property purposes

	PS0	PS1	PS2	PS3	"ANON"
Safe people *Can users be trusted to use data appropriately?*	– Research staff are required by contract to keep data confidential and to follow standard operating procedures for safe data collection – Research staff should have received the relevant privacy training. – Students/interns must sign confidentiality agreements and must follow institutional rules for how and where data will be stored after collection – Access rights should be limited to a few individuals who absolutely need to access the data – Documentation of who has access and what the access rights are should be maintained and updated regularly; temporary access should be revoked in a timely manner – Legal agreements should be established with any external parties who can access the data (such as processors or collaborators)	– Research staff are required by contract to keep data confidential and to follow standard operating procedures for safe data collection – Research staff should have received the relevant privacy training. – Students/interns must sign confidentiality agreements and must follow institutional rules for how and where data will be stored after collection – Documentation of who has access and what the access rights are should be maintained and updated regularly; temporary access should be revoked in a timely manner – Legal agreements should be established with any external parties who can access the data (such as processors or collaborators)	– Research staff are required by contract to keep data confidential and to follow standard operating procedures for safe data collection – Research staff should have received the relevant privacy training. – Students/interns must sign confidentiality agreements and must follow institutional rules for how and where data will be stored after collection – Documentation of who has access and what the access rights are should be maintained and updated regularly; temporary access should be revoked in a timely manner – Legal agreements should be established with any external parties who can access the data (such as processors or collaborators)	– Research staff are required by contract to keep data confidential and to follow standard operating procedures for safe data collection – Research staff should have received the relevant privacy training. – Students/interns must sign confidentiality agreements and must follow institutional rules for how and where data will be stored after collection – Documentation of who has access and what the access rights are should be maintained and updated regularly; temporary access should be revoked in a timely manner – Legal agreements should be established with any external parties who can access the data (such as processors or collaborators)	– Access, reading and writing rights of all internal research team members should be documented and regularly updated – Researchers should determine whether agreements need to be in place with external partners or with students for business, intellectual property or data ownership reasons; if not applicable, data may be shared freely and/or openly published – If data become re-identified in the future due to enhanced technological methods, every third party using the data is independently responsible for treating the data as personal data thereafter (i.e. it is not the original data collector's responsibility to inform or monitor secondary users of the data)

	PS0	PS1	PS2	PS3	"ANON"
Safe data *How to minimize disclosure risk within the data itself?*	Researchers should: – Determine if research goals can be completed without directly identifying data – Directly identifying information should be separated from indirectly identifying information, for example in a separate linking table/key file. – In some cases, it may be appropriate to mask the directly identifying information, e.g. via hashing.	Researchers should: – Determine if research goals can be completed without indirectly identifying data – Determine if research goals can be completed without specific variables that are the sole reason for re-identification, or by an alternative variable that is less identifying (e.g. age or year of birth instead of full birthdate). – Screen text fields for identifying information. – Generalize or remove unique data points/ extreme values. – Remove unnecessary identifying information. – Re-code data to a less identifiable form. – Use meaningless pseudonyms/-random ID numbers, whenever possible	Researchers should: – Determine if research goals can be completed without indirectly identifying data – Determine if research goals can be completed without specific variables that are the sole reason for re-identification, or by an alternative variable that is less identifying (e.g. age or year of birth instead of full birthdate). – Screen text fields for identifying information. – Generalize or remove unique data points/extreme values. – Remove unnecessary identifying information. – Re-code data to a less identifiable form.	Researchers should: – Determine if research goals can be completed without the use of a linking table/key file	Researchers should: Check with experts to confirm that data are in fact anonymous; choose experts appropriate to your discipline that can appropriately assess the type of data in question.

	PS0	PS1	PS2	PS3	"ANON"
Safe settings *How is unauthorized access prevented?*	– Institutions should develop faculty-level/ discipline specific standard operating procedures for safe data collection and storage – Research teams should establish data collection and storage protocols for all team members to follow to minimize privacy risks with the data collection – Data should be stored locally: only shared with external partners under strict circumstances, with secure methods for data transfer/sharing and with legal agreements in place between parties – Highest level of security for methods of collection and storage of data must be used. If subjects are vulnerable or the nature of the data is very sensitive/ potentially harmful to the subjects, additionally security measures beyond standard options may be necessary (e.g. additional encryption or use of air-gapped computers)	– Institutions should develop faculty-level/discipline specific standard operating procedures for safe data collection and storage – Research teams should establish data collection and storage protocols for all team members to follow to minimize privacy risks with the data collection – Data should be stored locally: only shared with external partners under strict circumstances, with secure methods for data transfer/sharing and with legal agreements in place between parties – In general, highest level of security for methods of collection and storage of data should be used, particularly when collecting data from vulnerable subjects or when the nature of the data is very sensitive/potentially harmful to the subjects. In some cases, a moderate level of security may be appropriate, if the	– Institutions should develop faculty-level/ discipline specific standard operating procedures for safe data collection and storage – Data should be stored locally, but may be shared with external partner as long as appropriately secure methods for data transfer/ sharing are used and legal agreements are in place between parties – Level of security for methods of collection and storage of data will vary depending on sensitivity of the collected data and vulnerability of the subjects. Security will range from moderate to the highest level; an appropriate level of security should be determined with the help of privacy and security experts – Data published in a third-party repository must only be accessible upon request and data must only be shared with external parties if secure methods are	– Data should be stored locally, but may be shared with external partner as long as appropriately secure methods for data transfer/sharing are used and legal agreements in place between parties – Level of security for methods of collection and storage of data will vary depending on sensitivity of the collected data and vulnerability of the subjects. Security requirements for this type of data are generally low, but an appropriate level of security should be determined with the help of privacy and security experts based on the nature of the data – Data may be openly published without data access restrictions only if the linking table/key file has been deleted; otherwise the data must only be accessible upon request and data must only be shared with external parties if secure methods are	– Data collection and storage methods should meet good data management standards, but privacy issues do not apply – Security issues may apply if the data contain confidential business or intellectual property information; these issues should be reviewed with security experts – Data may be openly archived and published without data access restrictions as long as no business confidentiality or intellectual property issues apply to the data – A trusted and discipline specific repository should be used for archiving – Published data must be licensed so that re-users know what they are allowed to do with the data

		vulnerability of the subjects or risk of harm is low, but this should be discussed with privacy and security experts Data published in a third-party repository must only be accessible upon request	used for data sharing and with legal agreements are in place between parties.	used for data sharing and with legal agreements are in place between parties.	

	PS0	PS1	PS2	PS3	"ANON"
Safe output *Is there a risk of disclosure in the statistical results (e.g. tables, figures)?*	– Screen output data for disclosure risk – Determine if results of research may have an impact on society/ individuals with characteristics similar to research participants. Discuss ethical concerns and consequences of research findings with ethics committee – Secondary users are responsible for screening output for re-identification – When assessing requests for access, consideration should be given to the impact on the participants	– Screen output data for disclosure risk – Determine if results of research may have an impact on society/ individuals with characteristics similar to research participants. Discuss ethical concerns and consequences of research findings with ethics committee – Secondary users are responsible for screening output for re-identification – When assessing requests for access, consideration should be given to the impact on the participants	– Screen output data for disclosure risk – Determine if results of research may have an impact on society/ individuals with characteristics similar to research participants. Discuss ethical concerns and consequences of research findings with ethics committee – Secondary users are responsible for screening output for re-identification – When assessing requests for access, consideration should be given to the impact on the participants	– Screen output data for disclosure risk – Determine if results of research may have an impact on society/individuals with characteristics similar to research participants. Discuss ethical concerns and consequences of research findings with ethics committee – The linking table/ key file and other ID variables should not be provided to secondary users, to avoid the risk of re-identification based on the output data – When assessing requests for access (if applicable), consideration should be given to the impact on the participants	– Determine if results of research may have an impact on society/ individuals with characteristics similar to research participants. Discuss ethical concerns and consequences of research findings with ethics committee – Prior to open publication of the data, consideration should be given to possible unforeseen impact that the anonymous/anonymized data could have on individuals. These issues can be discussed with privacy and ethics experts.

– Prior to open publication of the data, consideration should be given to possible unforeseen impact that the anonymous/anonymized data could have on individuals. These issues can be discussed with privacy and ethics experts.

COLOPHON

Risk Management for Research Data about people. *A generic matrix to be used by data stewards and research supporters for the assessment of privacy risks with research data and determination of appropriate methods for risk management.*

PUBLICATIION DATE | December 2019
DOI | 10.5281/Zenodo.3584333

LCRDM Task Group Anonymization
Jessica Hrudey (Free University/VU),
Jan Lucas van der Ploeg (University Medical Centre Groningen - UMCG),
Joan Schrijvers (Wageningen University & Research),
Arnold Verhoeven (University Maastricht),
Henk van den Hoogen (University Maastricht/liaison LCRDM advisory group),
Marjolein Sijbers-Klaver (University Medical Centre Utrecht – UMCU),
Santosh Ilamparuthi (TU Delft),
Toine Kuiper (TU Eindhoven),
Erik Tjong Kim Sang (eScience Center),
Niek van Ulzen (University of Applied Sciences Amsterdam/ HvA),
Yvonne Drost (Cultural Heritage Agency),
Ingeborg Verheul (LCRDM)

DESIGN | Nina Noordzij, Collage, Grou