Publishing and Sharing Sensitive Data

Key messages

- The advantages of publishing your sensitive data will probably far outweigh any potential disadvantages when simple and appropriate steps are taken
- Publishing your data, or just a description of your data (i.e., the metadata), means that others can discover it and cite it
- You can publish a description of your data without making the data itself openly accessible
- You can place conditions around access to published data
- Sensitive data that has been confidentialised can be shared

Author: Dr. Sarah Olesen, Australian National Data Service

About this guide

"Despite journal and funder policies requiring data sharing, there has been little practical guidance on how data should be shared" [1]

This Guide outlines best practice for the publication and sharing of sensitive research data in the Australian context. The Guide follows the sequence of steps that are necessary for publishing and sharing sensitive data, as outlined in the 'Publishing and Sharing Sensitive Data Decision Tree' (Figure 1, next page). It provides the detail and context to the steps in this Decision Tree. References for further reading are provided for those that are interested.

By following the sections below, and steps within, you will be able to make clear, lawful, and ethical decisions about sharing your data safely. It can be done in most cases!

How the Guide interacts with your institutional policies

This Guide is not intended to override institutional policies on data management or publication. Most researchers operate within the policies of their institution and/or funding arrangement and must, therefore, ensure their decisions about data publication align with these policies. This is particularly relevant for Intellectual Property, and sometimes, your classification of sensitive data (e.g., NSW Government Department of Environment & Heritage, Sensitive Data Species Policy) or selection of data repository. The Guide indicates the steps at which you should check your institutional policies.

Table of Contents

Publishing and Sharing Sensitive Data Decision Tree 4

Why publish and share sensitive data? 4

1. What are sensitive data? 6
   - 1.1 Defining sensitive 6
     - 1.1.1 Data about people 6
     - 1.1.2 Beyond human data 6
   - 1.2 An illustration of sensitive ecological data 7
   - 1.3 Sensitive by context 7
2. I have sensitive data – now what? 8
   - 2.1 Confidentialisation: A note on terminology 9
   - 2.2 What's legal? 10
3. Confidentialising human and personal data in 3 Steps 10
   - 3.1 Remove direct identifiers 10
   - 3.2 Consider removing or modifying indirect identifiers 11
     - 3.2.1 Methods of modifying data to limit identification 13
   - 3.3 Confidentialising sensitive ecological data 13
   - 3.4 Managing the risk of re-identification 14
   - 3.5 While you're at it; formatting data for publication 14
4. Ethical considerations 14
   - 4.1 For new human data, or when contact with research participants is possible 15
     - 4.1.1 What to include in a consent form requesting data publication and sharing 15
     - 4.1.2 Example sentences for consent forms to request data publication and sharing 16
   - 4.2 For existing data, when re-contact with research participants is *not* possible 17
   - 4.3 Sharing sensitive data that you did not collect 17
     - 4.3.1 Data citation 18
5. Making data discoverable 18
   - 5.1 Conditional access to data: What is it; how do I do it? 19
     - 5.1.1 An example of published metadata with conditional access to data: The Australian Longitudinal Study of Women's Health (ALSWH) 20
     - 5.1.2 Metadata record for the ALSWH 22
6. Licensing 22
7. Your right to publish 23
   - 7.1 How do I know if I own the copyright for the data? 23
8. Depositing your data 23
   - 8.1 Where should I publish my data? 23
9. Acknowledgments 24
10. More information 25
11. References 26

Publishing and Sharing Sensitive Data Decision Tree

1. Define

Do I have sensitive data?

- Collecting new data
  See: Data planning
- Yes
- No

2. Data collection

- Previously collected by you
- Collected by others
  See: Data Citation

3. Ownership

Do you have the right to publish?
See: Intellectual Property and data sharing

- Yes
- No
  Publish metadata only

4. Confidentialise

Can the data be made non-sensitive?

- Yes
  See: Confidentialise your data
  Formatting data for publication
- No
  Publish metadata only

5. Research ethics

Was consent for data sharing given by (any human) research participants?

- Yes
- No: Data many still be shared if it has been confidentialised.
  First, see: What's legal and Ethical considerations

6. Licensing

All published Australian data intended for reuse should have a licence. This explains to the re-user how the data should be used and attributed to the original data owner.
Go to: Licensing, ANDS licensing page and AusGOAL

7. Making data discoverable

Write a data description, i.e., metadata, for publication in a data repository.
See: How to select a repository

8. Publishing and sharing data

Publish metadata and data? Or metadata only with conditional access to the data itself?

- At minimum, metadata can be made publicly available.
- For confidentialised data, it is often appropriate to have *public metadata* and *'conditional access'* to the data itself.
- Also see: How to select a repository

Why publish and share sensitive data?

Box 1. Quick definitions

Data publication versus sharing: Publication occurs when data are made. This includes having a publically-available description of the data and access, or information about conditional access, to the data itself. Data sharing occurs when data are made available to others, but does not always accompany publication (e.g., when data are shared among colleagues but not publically discoverable or available).

Metadata: data about your data; i.e., a description of your data. A common example of metadata is a library catalogue record.

Repository: a place where metadata records and — often but not always — data are stored. Data repositories usually have an online portal where members of the public can search for and discover data. There are institutional-specific, discipline-specific, and general repositories.

Confidentialised data: when data has been modified to remove or reduce the risk that people or subjects of the data can be identified.

Sensitive data has often been excluded from discussions about data publication and sharing. It was believed that sharing sensitive data is not ethical or that it is too difficult to do safely. This opinion has changed with greater understanding and use of methods to 'de-sensitise' (i.e., confidentialise) data; that is, modify the data to remove information so that participants or subjects are no longer identifiable, and the capacity to grant 'conditional access' to data. Requirements of publishers and funding bodies for researchers to publish and share their data have also seen sensitive data sharing increase[2-4].

Australia and other nations have significant and high-quality datasets that contain potentially-sensitive information. This includes epidemiological surveys of health, medical trial data, and ecological studies of endangered species. For reasons of finance, efficiency, conservation, and participant fatigue and disturbance, these can be reused rather than repeated[5, 6]. And in most cases, this can be done by removing or modifying aspects of the dataset that make its subjects identifiable (see Confidentialise your data) or attaching conditions to data access and reuse.

The advantages to sharing data go beyond meeting publisher or possible funder requirements. The benefits to the researcher and institution are clear. If data or a description of a dataset are published they are discoverable by others, and can thus be cited in favour of the original data collector or owner. This is the primary goal of data publication. There is also evidence that scholarly papers that are accompanied by published data are cited more often than those without. New collaborations and publications may eventuate if you choose to share your data with others for reuse. And finally, storing your data in a public repository affords secure and ongoing storage that may not be available in your current or changing research environment.

1. What are sensitive data?

1.1 Defining sensitive

Sensitive data are data that can be used to identify an individual, species, object, process, or location that introduces a risk of discrimination, harm, or unwanted attention. Under law and the research ethics governance of most institutions, sensitive data cannot typically be shared *in this form*, with few exceptions.

1.1.1 Data about people

Sensitive human data most commonly refers to sensitive *personal information*. That is, information that can be used to identify a person or group of people[7, 8]. Personal information can thus be thought of as one type of sensitive data. In most cases, personal information is sensitive when it directly identifies a person (e.g., name, Date of Birth, address) and accompanies one or more pieces of information from Table 1.

Specialised information about sharing data from or about Aboriginal and Torres Strait peoples and practices can be found in the AIATSIS Collection Access and Use Policy. You may also wish to consult the AIATSIS Guidelines for Ethical Research in Australian Indigenous Studies.

Specialised information about the use and management of data about Defence operations can be found via the Attorney-General's Department[9].

Table 1. Types of personal information, as defined by *Privacy Act 1988*, s 6(1)
racial or ethnic origin
political opinions
membership of a political association
religious beliefs or affiliations
philosophical beliefs
membership of a professional or trade association
membership of a trade union
sexual preferences or practices
criminal record
health and genetic information

1.1.2 Beyond human data

Sensitive data can also include data that reveals the location of rare, endangered or commercially-valuable species, or other conservation efforts[10]. The need for an agreed definition of sensitive

environmental and biodiversity data is well documented[10, 11]. The lack of clarity surrounding this definition is, at least in part, due to the sometimes-transitional nature of data sensitivity in this field. For example, a population of frogs may be small and declining in one location but not another, and thus, considered sensitive in the former but not latter.

Chapman and Grafton (2008)[11] at the Global Biodiversity Information Facility (GBIF) define sensitive data as information that *'if released to the public, would result in an 'adverse effect''* to the species or conservation activity. Particularly when the publication or sharing of this data may increase that risk of harm and/or result in an adverse effect on the species. This GBIF report further recommends that data owners clarify *which elements* of the dataset trigger this sensitivity, and clearly document that decision in the metadata for the dataset.

1.2 An illustration of sensitive ecological data

A survey was conducted in an area of New South Wales over a six-month period to estimate the population and breeding habits of several bird species. The data resulting from this survey included georeferencing data and breeding status of each species. One of the species included in this survey was the Glossy Black Cockatoo, which is listed as a 'vulnerable' species by the NSW Department of Environment & Heritage in their Sensitive Species List. This list states that the reason for the cockatoo's vulnerable status is *'risk of egg collection and nest disturbance'*.

Identifying the causes of a species' 'sensitivity' also enables the data owner to isolate which elements of the dataset may be removed or modified to confidentialise the dataset (Section 4.3).

Check: Are your data sensitive? If your data contains information from Table 1 *or* information about secret or sacred practices, *or* information that would result in an adverse effect on a species if made public, it is likely to be sensitive.

In addition to looking at the bare content of the data, ask whether, if shared, the data could *potentially cause harm or contribute to discrimination* to determine if the data are sensitive. This latter aspect relates to ethical considerations about data publication (see Section 4).

1.3 Sensitive by context

Some data are born sensitive, some achieve sensitivity, and some have sensitivity thrust upon them!

What is considered to be sensitive may differ with time and across groups of people or subjects. Non-sensitive data may become sensitive by their context, or when more information is added. The names of an organisation's or publication's subscriber list are not usually sensitive. But this list may be sensitive if the special interest of that group could expose subscribers to discrimination[12]. For example, an extreme or unpopular political affiliation. For ecological data, a species may be at risk of harm from

human activity in one geographic area but not another. In this latter case, sensitivity might be determined by the location that the data refers to rather than the species itself.

There are instances where data that are not obviously sensitive (i.e., does not include names or dates of birth), or has been confidentialised, may become sensitive again when context changes. Two common examples are:

1. Triangulation: When the identity or sensitivity of a participant or subject can be determined by combining several pieces of non-sensitive information. For example, with human data, if you have information about a person's age, occupation and family composition, it may not be difficult to identify them in a relatively small sample.

2. Data Linkage is bringing together two or more datasets that include the same person or subject of research. Alone, a dataset may not contain enough information to identify individuals or make place subjects at risk, but when combined with two or more datasets, this may now be achievable.

For example, Dataset A describes the medical history of a group of non-identifiable patients with a cognitive disorder. Dataset B includes employment information and public transport usage in the same, moderately small population. When these datasets are linked, there could be sufficient information about where patients work and live such that they may be identified.

An example from ecological research: Dataset A describes a species of falcon that is vulnerable to egg collection, and its population and location over time. Dataset B includes breeding patterns of the same falcon species over a similar period of time. When linked, there may be enough information to determine the location of falcons at nesting time.

Data owners and managers should always consider the possibility triangulation in their dataset and check for this. It is good practice to re-consider triangulation with the introduction of new data; that is, following data linkage. Section 3 of this guide shows you how to confidentialise a dataset so that participants or subjects cannot be identified, including in the context of triangulation.

2. I have sensitive data – now what?

How to publish and share sensitive data

The previous section describes sensitive data in its original form. In the following sections on Confidentialisation and Licensing we discuss ways you can modify original data so that it is *no longer sensitive*. This kind of modified data is still highly valuable and re-usable to some other researchers without placing research participants or subjects at risk. And the data can be published and shared with fewer legal and ethical restraints.

In most instances, confidentialised sensitive data is published with conditional access. This means that a description (i.e., 'metadata') is published in a data repository and its discoverability is not restricted – i.e., anyone can find and read a description of the data. And the data itself is accessible once some conditions (set by the researcher and/or the repository) are met. The public description of the data

includes this information about access and a link to how/where to apply for access. Common conditions of access to confidentialised data are set out below.

2.1 Confidentialisation: A note on terminology

The terms 'confidentialisation', 'de-identification', and 'anonymisation' are often used interchangeably. They can, however, also refer to somewhat different methods. *Confidentialisation* involves removing or altering data so that people or the subjects of the data cannot be identified. In the strictest sense, *de-identification* is simply the removal of identifying information. It is also typically used to refer to human data only. Finally, anonymisation typically refers to de-identified data, but is often used to describe confidentialised data also.

If the above terms are used in research documents, particularly information provided to research participants; it is recommended that the authors of these documents define their meaning of the terms to avoid confusion[5].

In this guide, we refer to confidentialisation and include non-human data in this definition. In line with the methods recommended by the Australian Bureau of Statistics, National Statistical Service, this includes:

1. **De-identifying the data**; that is, removing information that can directly or indirectly identify a person, persons, species, or other subject of research; *and*
2. **Continuing to manage the risk** of identification even after the dataset has been de-identified.

It is notable that the *Privacy Act 1988* s *6(1)* definition of de-identification[footnote1] is actually akin to the above definition of confidentialisation[13]. It is also notable that the *National Statement on Ethical Conduct in Human Research (2007)*[5] instead uses the terms 'identifiable', 're-identifiable', and 'non-identifiable' data.

[1] The Australian Government's Office of the Australian Information Commissioner defines de-identification in relation to the Privacy Act 1998 as:

1. *"removing personal identifiers, such as an individual's name, address, date of birth or other identifying information, and*
2. *removing or altering other information that may allow an individual to be identified, for example, because of a rare characteristic of the individual, or a combination of unique or remarkable characteristics that enable identification."*

2.2 What's legal?

Under the Australian *Privacy Act 1988*[7], sensitive human and personal data cannot generally be shared in their original form. However, once confidentialised, these modified data no longer trigger the Act. In other words, confidentialised sensitive data can legally be shared if the method of confidentialisation meets the standards of the *Privacy Act 1988*. The definition of confidentialisation (above) used in this Guide meets this standard. The steps in the next section show you how to confidentialise your data according to this standard.

It is worth noting that whilst the *Privacy Act 1988* does not apply to confidentialised data, it does apply to the *activity* of confidentialising the data (i.e., removing identifying information from the original, sensitive dataset). This activity is, however, explicitly condoned in the *Australian Privacy Principles* of the *Privacy Act 1988* as one of few exceptions to sensitive data use. This is because confidentialisation is considered a '*normal. .. practice'* that *'an individual may reasonably expect their personal information to be used or disclosed for'* without requiring specific consent[14]. For more information about human consent for data sharing see our section on 'What to include in a consent form'.

Key points on the legality of sharing sensitive data:

- Sensitive data cannot be published and shared in its original form in almost all cases
- Confidentialised data is no longer sensitive and can be shared
- The process of confidentialising human and personal data is allowed under the Australian *Privacy Act 1988*[14]

3. Confidentialising human and personal data in 3 Steps

The goal of confidentialisation for data sharing is to prevent participants or subjects of research (e.g., animal or plant species) from being identified and placed at risk of harm or discrimination. This involves removing or modifying information in the original (sensitive) dataset. The exact information that needs to be removed or modified will vary depending on the contents of the dataset and the reason that the data has been deemed sensitive.

The following steps were developed following review of key research papers and existing guidelines of relevant local and international authorities on data management.

3.1 Remove direct identifiers

Remove **all** direct identifiers from datasets intended for publication. A list of direct identifiers is shown in Table 2.

Table 2. Direct identifiers
Name
Initials
Address, including full or partial postal code
Spatial location *(e.g., latitude and longitude units with enough precision to potentially locate the subject)*
Telephone or fax numbers or contact information
Electronic mail addresses
Vehicle identifiers
Medical device identifiers
Web or internet protocol addresses
Biometric data
Facial photograph or comparable image
(*unconfidentialised*) Audiotapes
Names of relatives
Dates related to an individual (including date-of-birth)

Notes on Table 2: This list is produced from a review of recommended confidentialisation procedures for clinical health data conducted by Hrynaszkiewicz et al.[1] and identifying aspects of sensitive biodiversity data as defined in the Atlas of Living Australia's Sensitive Data Report[10]. These recommendations are consistent with those provided by other institutions (e.g., National Statistical Service). Information in italics was added by the authors of this guide to provide further clarification or broaden use to other forms of sensitive data.

3.2 Consider removing or modifying indirect identifiers

Datasets that contain **two or more of the indirect identifiers in Table 2** may identify participants when these identifiers are considered together (i.e., akin to triangulation). When two or more indirect identifiers are present, we recommend removing or modifying one or more of the indirect identifiers until the risk of identification is negligible. If you are still unsure if the risk of identification is negligible or not, we recommend seeking advice from an independent researcher, data manager, or member of a research ethics committee.

Table 3. Indirect identifiers
Place/location of treatment, education, service use
Name of professional or business/service responsible for healthcare, education, service
Sex
Rare disease, condition, experience, treatment, or other characteristic
Risky behaviours (e.g., Illicit drug use)
Place of birth
Socioeconomic data, such as occupation or place of work, income, or education level
Household and family composition
Body measures (e.g., height, weight)
Multiple pregnancies Ethnicity Indigenous status
Year of birth or age Verbatim responses or transcripts Dates of sensitive events Small cell sizes — i.e., when the number of subjects with the characteristic is small
Notes on Table 3: This largely includes information from a review of recommended confidentialisation procedures for clinical health data conducted by Hrynaszkiewicz et al.[1] This original list has been modified to expand its relevance to non-medical data. We have also included Indigenous status as a separate identifier to Ethnicity for the Australian context because this population is often oversampled and/or living in small or specific communities.

Removing versus modifying: Completely removing information from a dataset ensures this information cannot be used to identify participants or subjects. Sometimes, however, information can be modified enough that it no longer poses risk of identification and can thus remain in the dataset. This involves more effort but is a good option if complete removal of the information de-values the dataset.

'Sort of sensitive' – Grey areas

You may choose to remove or amend other aspects of your dataset before publishing and sharing your data if they are not considered essential to the dataset, or if their presence will impede sharing. Financial information about an individual, household, or organisation is a common example of this. It is not considered 'sensitive' under the *Privacy Act 1988*; however, individuals typically do not want their financial dealings to be shared and traced back to them.

3.2.1 Methods of modifying data to limit identification

- Combining responses into categories, or a fewer number of categories than in the original dataset. This is a good option if only a small number of people or subjects possess the characteristic. For example, year of birth can be collapsed into 5- or 10-year age bands if only a few people in a dataset report a specific birth year.
- Top and bottom coding: collapsing categories into upper and/or lower thresholds. This is a good option if only a small number of people have high or low measurements on a characteristic. For example, if few people report that they have more than five children, these participants can be combined with those who report five children and recoded as '5+ children'.
- Rounding dates, times, or measurements reduces the risk of identification when only a small number of people or subjects have a specific value.
- Cell suppression: involves creating 'missing data' cells if the inclusion of this data poses a risk to identification. Single cells may be deleted, or all data for an at-risk participant or subject (for more information on suppression)
- For further information on more complex methods of modification, including data rounding and re-sampling, see National Statistical Service.

3.3 Confidentialising sensitive ecological data

Methods for confidentialising environmental and biodiversity data are less established than those for human data. For human data, sensitivity usually relates to the identification of individuals. Thus, directly identifying information must be removed for confidentialisation. However, the sensitivity of ecological data typically does not typically relate to the naming of the species itself, but to accompanying information about its location and/or dates for breeding, fruiting, or migration. For example, the illustrative dataset in Section 2.2 is not sensitive simply because it includes data from the Glossy Black Cockatoo, but because it is accompanied by data about its precise location at breeding time.

Recommended steps to confidentialisation:

1. Identify and record which elements of your ecological dataset make it sensitive[11]. This involves first determining why the species or location is at risk, then identifying what data in the dataset provides that information.

   Using the illustration in Section 2.2, the data causing sensitivity (i.e., placing the species at risk) in this dataset is about breeding status because the species' 'vulnerable' status is due to risk of egg collection and nest disturbance.

2. Remove or modify enough information so that the dataset ceases to be sensitive whilst retaining as much detail as possible for potential reusers.

   In the example from Section 2.2, information about breeding status could be removed from the dataset to alleviate sensitivity. Alternatively, the precision of georeferenceing data could be reduced so that locating bird nests was not possible. Such processes of data modification are akin to the methods of described above for human data.

3.4 Managing the risk of re-identification

Once data has been confidentialised, the risk of re-identification must be reviewed when these data are linked with other data. Data linkage is the merging of two or more separate datasets that contain data from the same people or subjects. It is becoming increasingly common in epidemiology, medical, social and ecological sciences because it enables researchers to understand people's or subjects' context in more detail by adding more information without having to collect new data. Data linkage also derives greater value from existing datasets.

Like triangulation, data linkage can mean that confidentialised participants or subjects can become re-identified because one or more pieces of potentially-identifiable information have been added (by the newly linked data). This possibility must be assessed when the data are linked by treating the new, linked dataset as a non-confidentialised dataset and following the steps above.

3.5 While you're at it; formatting data for publication

The dataset you submit to a repository for publication should be[1]:

- Cleaned; checked for errors, outliers, duplicates, and missing data (and genuinely missing data should be annotated as such)
- Annotated; variables and objects (and their categories if relevant) should be labelled clearly and a key provided if necessary. If variables or objects are modified from their original form during confidentialisation, note this.
- In a format that is open, and easy to transform and archive. Strive to use a data format that is readable and malleable in a variety of commonly-used operating systems and programs. Non-proprietary ('open') formats are also recommended to enhance accessibility. Information about the data format should be provided in the metadata record. If a specialised program is required to read and analyse the data, then this should be provided alongside the data where possible.

Check your chosen data repositories for any specific requirements about data formatting.

4. Ethical considerations

In addition to meeting legal standards, researchers have ethical obligations towards participants and research subjects. These include preserving privacy and avoiding any possible harm arising from participation in research and its subsequent publication. The ethical management of data must be the primary concern of researchers to maintain participants' trust and research integrity.

Check: Are your data sensitive? In addition to looking at the bare content of the data, ask whether, if shared, the data could *potentially cause harm or contribute to discrimination?* to determine if the data are sensitive.

4.1 For new human data, or when contact with research participants is possible

Obtaining consent for data publication and sharing from research participants *before* the data are collected optimal. This is not only best practice, but avoids the expense, delay and loss of use of data that may be occasioned by attempting to obtain consents later in the research process. Concerns that participants will refuse to participate in research if data sharing is requested are likely to be unfounded[15, 16].

The request for participants' consent to publish and share their confidentialised data should involve:

1. Including information about the processes of data confidentialisation, publication, and sharing in the 'information sheet' that is provided to people before they agree to participate in the research study and before they are asked to consent to data collection. This information must understandable to the participant and have sufficient detail such that they can make an informed decision about their consent to publish and/or share the data they provide. Participant information sheets must be approved by your Human Research Ethics Committee/s (HREC) before any aspect of the study commences. It should (briefly) cover: procedures for maintaining confidentiality, data publication, and the conditions under which data sharing can occur, including whether there is a prospect of sharing the data with researchers outside those conducting the study[5].
2. Specifically requesting consent for data sharing and/or publication in the 'consent form', which must also be approved by your HRECs before research begins. Example wording to request data publication and sharing in consent forms are provided below.
3. Familiarise yourself with any requirements for information sheets and consent forms in your jurisdiction with regards to data publication and sharing. For example, whilst the current National Ethics Application Form (NEAF; required for all Australian research involving human participants) does not include questions about data publication and sharing, research conducted in hospitals in the State of Victoria must also complete a Victorian-Specific Module. This Module (or form) *does* refer to data storage and future research.

4.1.1 What to include in a consent form requesting data publication and sharing

Statements about data publication and sharing in participant consent forms should:

1. Avoid precluding data confidentialisation, publication, and sharing[17]
2. State the possibility of future data publication (including storage in a repository) and sharing
3. State the conditions under which access to the data may be granted to others. This may include the process of confidentialisation and possibility other conditions, such as approval by the original research team (see later section on conditional access). In some cases, it may also be appropriate to provide an opportunity for participants to select whom they agree to share their data with (and

whom they don't).

4. Be documented with the collected data so subsequent users of the data are aware of the conditions agreed to by participants[5].

4.1.2 Example sentences for consent forms to request data publication and sharing

Where data is intended to be public or accessed with little restriction:

'The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential..'[18]

'Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public'[17]

'I agree that research data gathered for the study may be published provided my name or other identifying information is not used'[19]

For data that will have conditional access:

'..other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form'[20]

The above example may be adapted to include specific access conditions that you intend to apply to its reuse:

'*other genuine researchers may request access to confidentialised data in the future. Access will only be granted if they agree to preserve the confidentiality of the information as requested in this form. Their access will also require approval from the original research team'*

You may also consider giving participants the opportunity to select whom they agree to share their data with (and whom they don't). For example, from a list of likely data re-users.

4.2 For existing data, when re-contact with research participants is *not* possible

As stated above in the sections on What is sensitive data?, What's Legal? and Confidentialising your data, sensitive data can legally be shared without explicit consent from research participants if[5]:

1. The information given to participants prior to their consent for data collection indicated future use of the data[footnote2], **OR**
2. The opportunity to gain consent no longer exists or is not practical, and
3. The data have been properly confidentialised, and
4. The process of confidentialisation matches the definition provided in the *Privacy Act 1988*, and
5. There is no risk that publishing or sharing the data will cause harm or contribute to discrimination towards the research participants or subjects, and
6. Information Sheets and Consent forms from the original data collection did not *preclude* sharing.

Recommendation: If you are unsure whether you meet these requirements, it is recommended that you *seek the opinion of your local HRECs* and colleague/s outside your research project.

4.3 Sharing sensitive data that you did not collect

What do you do about data publication and sharing if you are not the data owner? That is, you are the data re-user or 'secondary data user'.

Data reuse is already relatively widespread in some research disciplines that deal in confidentialised data, in fields such as epidemiology. The question above is becoming increasingly common in response to requirements of journals to publish data alongside scholarly articles.

In most instances, you cannot publish data that you did not collect because you do not own the Copyright for that data. Exceptions are data that have been licensed by the data owner to allow redistribution. (e.g., see AusGOAL's endorsed Creative Commons licenses.)

A note on research ethics: Like original researchers, secondary data users have an '*obligation to ensure that the data are used responsibly and respectfully, and that the privacy of participants is safeguarded*'[5]. Thus, data re-users must uphold any conditions of data use that were specified to the participants by

[2] In cases where participant Consent Forms did not refer specifically to data publication or sharing (though not precluded it either) and Information Sheets did, consent to participate in the project itself allows sharing. This is because consent implies an understanding and agreement to the Information Sheet.

the original researchers, as well as those conditions of re-use outlined by the data owner or manager when access for re-use was granted.

4.3.1 Data citation

If you are a data re-user, you should also reference or cite the original source of the data in all the articles, presentations, and grant applications based on that data. This enables the data owner to track the use of their data, as well as linking your work with other articles based on this dataset. See the ANDS Guide to Data Citation for detailed information on how to do this.

Recommendations:

- *If the data you are re-using are licensed*, follow the conditions of that license regarding data sharing (typically termed 'redistribution'). If redistribution is allowed, you may be able to share the data and attribute to the data owner. Alternatively, cite the data source in any public description of your research so that others can also request access.
- If the data you are re-using is not licensed, contact the data owner or manager for instructions about publication and sharing.

Whenever you are reusing data, you should cite the original source in *all* scholarly outputs.

5. Making data discoverable

A public metadata record in a repository is the best and easiest way to make your data discoverable by others. This is a description of your data in a public catalogue. Your chosen repository will provide instructions about what metadata are required before publication. This will almost always be information that you already have in grant applications, project reports, or articles. A public metadata record allows others to find and cite your data. If the data themselves are not published alongside the metadata record (most often via a link in the record, see example below), the record also provides the potential re-user with information about how the data can be accessed.

You can publish a description of your data without making the data themselves freely available

It is common for researchers with sensitive data to want to publish and share their data but still hold concerns about open publication or access. This is largely amenable by publishing the data with conditional access.

5.1 Conditional access to data: What is it; how do I do it?

Conditional access occurs when a metadata record is available to the public (i.e., published in a public repository) but access to the data themselves occurs only after pre-determined conditions are met. These conditions are set by the researcher or data owner, and/or the data repository. They may include requiring the potential data re-user to:

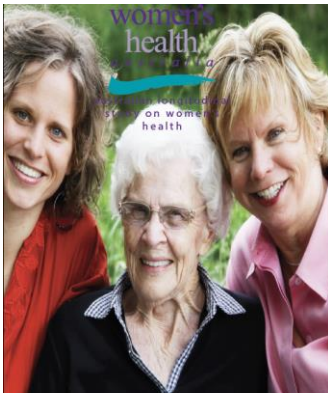
- Register and/or provide contact details
- Provide information about how they will use, store, or manage the data
- Agreed to conditions of data security, privacy
- Agree that they may be contacted by the data owner for purposes of collaboration or otherwise
- Pay an access fee
- Meet other conditions included in Consent Forms and Information Sheets agreed to by original (human) research participants[5]

Reasons you might want to make your data discoverable but with conditional access:

- To ensure re-users are genuine researchers
- To ensure re-users are aware and agree to maintain confidentiality or secure storage of the data
- The data (or some of the data) are under embargo;
- You would like to maintain some oversight over *who* uses the data
- Or for what *purpose* they are using the data
- You would like to be informed about who uses the data so you can collaborate
- You specified conditional access in the participant consent form or Research Ethics Application form
- Others can use the data for new and valuable purposes. Ideally, research data should be widely disseminated —whether or not the data creator or manager considers it valuable. It is impossible to know what applications of the data may exist in the future!

5.1.1 An example of published metadata with conditional access to data: The Australian Longitudinal Study of Women's Health (ALSWH)

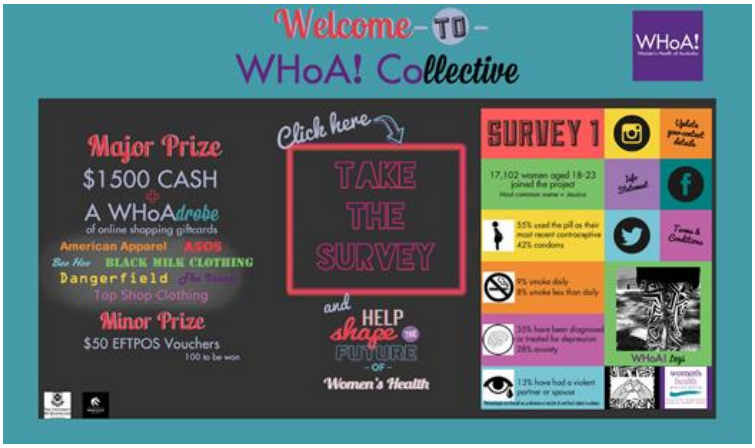
Sensitive data sharing: Benefiting women's health



"Data sharing is fundamental for ALSWH as a national research resource to strengthen the evidence base for supporting development of women's health policy and practice. We are fully committed to making our data available and encouraging collaboration between researchers in wide-ranging fields. Data sharing among multidisciplinary groups provides the opportunity for fresh perspectives and for gaining new insights and knowledge on women health."
Professor Gita Mishra, ALSWH Director

The *Australian Longitudinal Study on Women's Health* (ALSWH) is a collaborative project of The University of Newcastle and The University of Queensland and has been funded by the Australian Government for almost two decades. Since 1995 over 50,000 women have been surveyed, most choosing to remain part of the study for many years. ALSWH collects information about changes in the mental, physical, and social health of everyday women and their families over time. It also gathers data about life events, employment, and health service use. This level of detail and longevity is rare and highly valuable on an international scale. ALSWH has enabled data users to produce a rich and accurate portrait of women's health and the experiences that benefit or hinder their wellbeing.

If you haven't heard of ALSWH, you've probably read a report, paper, or news article based its data. You may even use a public service or program guided by its findings. Data from ALSWH has led to almost 500 peer-reviewed papers across domains of clinical medicine, public health, and ageing research. These include highly-cited studies that, for example, identify factors contributing to obesity in Australian women[1] and demonstrate the link between increased physical activity and reduced symptoms of depression[2]. ALSWH also contributes directly to national health policies by informing recommendations for services and programs for chronic health conditions (e.g., diabetes), interpersonal violence, carers, nutrition and physical activity.



ALSWH adds considerable value to other data sources by supporting sub-studies and data linkage projects. With participants' consent, ALSWH data has been linked to women's Medicare data to reveal the uptake and use of new and sometimes-controversial health service schemes[3], and evaluate the costs of novel healthcare programs[4]. The NHMRC-funded 'Mother's and their Children's Health' (MatCH) sub-study will collect data for children from a sample of mothers already participating in the ALSWH.

This represents an unrivalled opportunity to consider the links between children's wellbeing and the full range of health, social, and service-use information already collected from their mothers at earlier stages of the study.

The substantial contribution of the ALSWH to scientific knowledge, public policy, and population health is well evidenced by its international reputation and continued Government funding. But **did you know that ALSWH is also a leading example of sensitive data sharing?** The survey collects information about participants' health and lifestyle but is used by over 650 collaborators — most of whom are not part of the original research team.

How they do it

The ALSWH has always had clear policies for maintaining the quality and security of data and enabling others to find, request and receive conditional access to the data[5,6]. These policies are founded on the research team's belief that this dataset is a 'public resource'. ALSWH data-sharing practices address legal and ethical considerations whilst promoting re-use, collaboration, and research integrity.

Discoverability: ALSWH has a public website with detailed information about the study and data accessibility. Metadata records for the study also exist in several national repositories, including Research Data Australia, Australian Data Archives (where ALSWH data is also archived), and Trove.

Conditional access to data for re-use: Potential re-users must complete an 'Expression of Interest' form via the ALSWH project page. The re-user must provide information about themselves and a description and justification for their intended use of the data. This process is designed to maintain study integrity and prevent research overlap. The application is reviewed by the 'ALSWH Publications, Substudies and Analyses (PSA) committee'. If approved, the confidentialised data is received by the data re-user.

Legal & Ethics: Directly identifying information is removed from all ALSWH datasets. This information is kept at a separate, secure location to the datasets and is not accessible. All users of ALSWH data are required to sign a Privacy Protocol similar to a data license, which specifies that the data cannot be distributed or used for research purposes beyond those specified in the 'Expression of Interest' form. The ALSWH Information Statement to participants (which requires participant agreement before they enter the study) does not preclude data sharing. Participants are further informed that *'my answers will not be linked with my personal details and so will not be identifiable'.*

Authors: Dr. Sarah Olesen (ANDS), Prof. Gita Mishra, & A/Prof. Leigh Tooth

References

Ball K, Brown W, Crawford D. Who does not gain weight? Prevalence and predictors of weight maintenance in young women. *International Journal of Obesity*, 2002, 26(12), 1570-1578.

[2] Brown W. Prospective study of physical activity and depressive symptoms in middle-aged women. *American Journal of Preventive Medicine,* 2005, 9(4), 265-272. DOI: 10.1016/j.amepre.2005.06.009, PMID: 16242588

[3] Byles JE, Dolja-Gore X, Loxton D, Parkinson L, Stewart Williams J. Women's uptake of Medicare Benefits Schedule mental health items for general practitioners, psychologists and other allied mental health professionals. *Medical Journal of Australia*, 2011, 194(4), 175

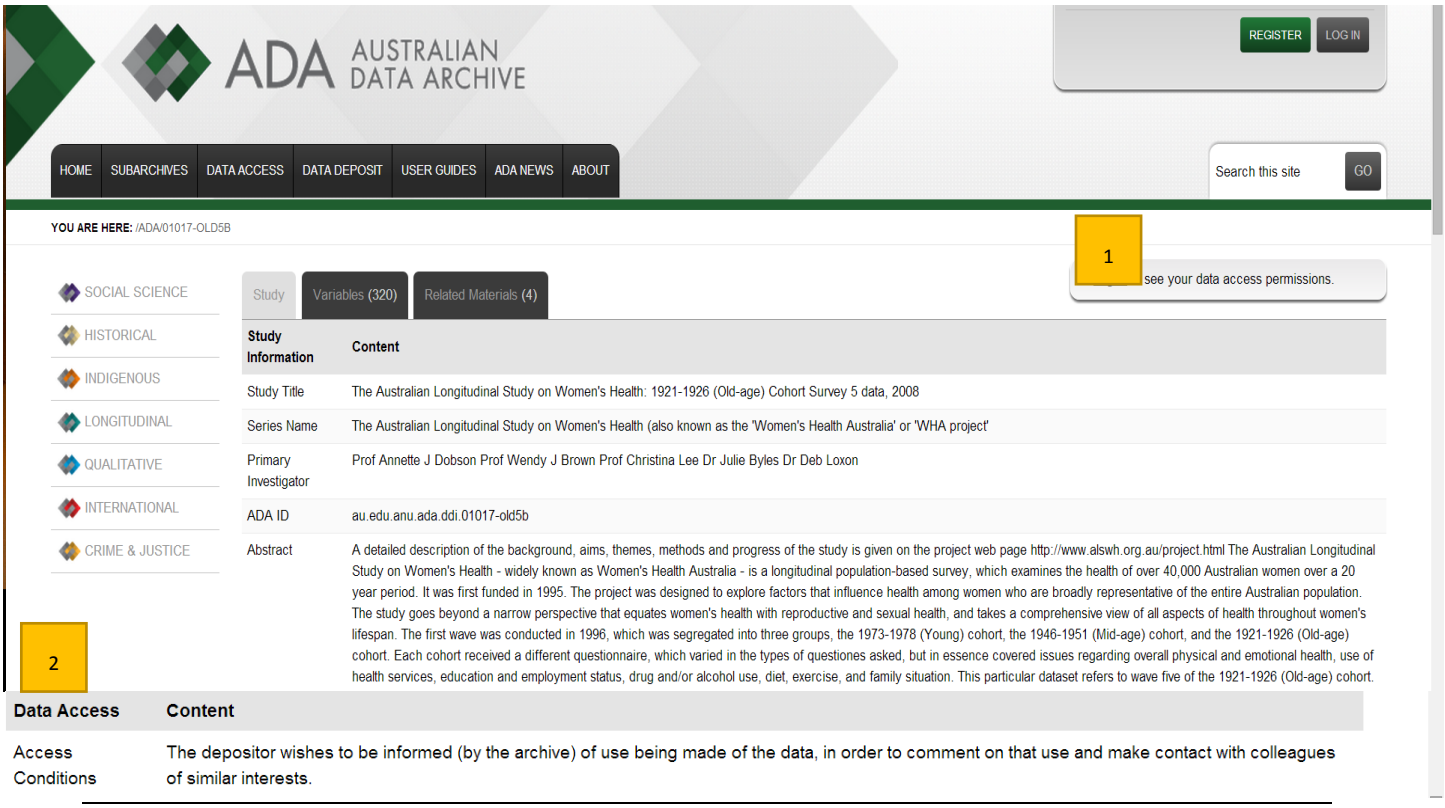
[4] Lowe J, Byles J, Dolja-Gore X, Young A. Does systematically organized care improve outcomes for women with diabetes. *Journal of Evaluation in Clinical Practice*, 2010, 16(5), 887-894

[5] http://www.alswh.org.au/for-researchers

[6] Australian Bureau of Statistics, 2009, *ABS Data Quality Framework*, cat. no. 1520.0, http://www.abs.gov.au/

5.1.2 Metadata record for the ALSWH

This is what the metadata record for the archived Australian Longitudinal Study on Women's Health data in the Australian Data Archive (data repository). In addition to requiring a login 1 to access data in the Australian Data Archive (i.e., you must register and be approved by the repository), the 'Access Conditions' 2 for this dataset are clearly explained.



6. Licensing

All Australian data intended for reuse should have a licence. This includes your confidentialised dataset. A licence is a document that clearly sets out how the data can be used and attributed to the original data owner. Without a licence, it is unclear how your data can be reused and this may discourage the potential re-user.

Licences come in varied forms, ranging from few to many restrictions on reuse. Some data repositories have their own licensing documents (e.g., Australian Data Archive's 'Access Categories'). Others require open access; that is, non-restricted access (e.g., Dryad).

AusGOAL provides a suite of endorsed licenses (ranging from few to more conditions of reuse) and a

template for a Restrictive License. On the AusGOAL website you can select and follow simple steps to apply one of their licenses to your data, or design a Restrictive Licence to suit your purposes. AusGOAL is endorsed by ANDS and the Australian Governments'.

It is important to note that applying an open license to your data does not allow you to publish sensitive data, or act as a substitution for data confidentialisation. Sensitive data remains sensitive even with a license, and thus cannot be published without participant consent. Accordingly, this Guide recommends that you apply a licence, publish, and share your data *after* it has been confidentialised as described above.

For more detail see the ANDS webpage on Data Re-use and Licensing Frameworks, and the ANDS Guide to AusGOAL.

7. Your right to publish

In most cases, the person or institution that publishes the data must also hold the appropriate rights (such as copyright) to do so. An AusGOAL endorsed licence can only be applied by to the rightsholder of that data.

7.1 How do I know if I own the copyright for the data?

Ownership of copyright and instances of copyright waiver differ across Australian Institutions. As copyright is an aspect of intellectual property, we recommend that you look up the Intellectual Property Policy of your institution or employer. If still unclear, seek advice from your Research Office or Research Services Division.

For more detailed information: ANDS Guide to Copyright, Data and Licensing.

8. Depositing your data

8.1 Where should I publish my data?

There are many data repositories to choose from. Some repositories provide a catalogue of metadata only and link to, or reference, the storage location of the data (e.g., Research Data Australia). Others catalogue metadata *and* store the data themselves (e.g., Australian Data Archive, Figshare). Repositories can also be institution-specific, discipline- or content-specific, or general. You can find a list

of international repositories at databib.org.

When selecting a place to publish your data, consider:

- Whether your institutional, employer, funder, or publisher mandates or recommends a particular repository
- Whether your research discipline has conventions around where to publish data
- Whether you want to publish your metadata and data in the same place
- What metadata the repository requires
- The format in which the repository requires your data
- Any financial costs incurred
- Whether the repository enables tracking of data citations by allocating your data a unique identifier (see DOIs)
- Whether and how the repositories manage conditional access to data, and whether conditional access is managed by the repository or the data owner.

9. Acknowledgments

The author acknowledges and thanks Dr. Greg Laughlin (Principal Policy Adviser, ANDS), Baden Appleyard (National Programme Director, AusGOAL), Professor Michael Martin (Research School of Finance, Actuarial Studies and Applied Statistics, The Australian National University; Chair, Humanities and Social Sciences Delegated Ethical Research Committee; Chair, Science and Medical Delegated Ethical Research Committee, The Australian National University), and ……. For their invaluable advice and comments during the writing of this Guide.

10. More information

Relevant National Guidelines

- National Health and Medical Research Council, Australian Research Council, the Australian Vice-Chancellors' Committee, National Statement on Ethical Conduct in Human Research (Updated March 2014)

Defining sensitive data

- Atlas of Living Australia, 'Our secrets are not your secrets: Sensitive data report'
- Global Biodiversity Information Facility, Guide to Best Practices for Generalising Sensitive Species Occurrence Data
- NSW Government Department of Environment, Climate Change and Water, Sensitive Species Data Policy

Preparing data for deposit

- UK Data Service 'Depositing Sharable Survey Data'

Related international guides

- UK Data Service 'Depositing Sharable Survey Data'
- Digital Curation Centre (UK), How to Develop a Data Management and Sharing Plan
- Inter-university Consortium for Political and Social Research (ICPSR), Guide to Social Science Data Preparation and Archiving

11. References

1. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ. 2010;340:c181.
2. National Institutes of Health. Final NIH Statement on Sharing Research Data. National Institutes of Health, Office of Extramural Research. Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.
3. PLoS ONE. PLOS Editorial and Publishing Policies. PLOS. Available from: http://www.plosone.org/static/policies#sharing.
4. Wellcome Trust. Policy on data management and sharing. Wellcome Trust. Available from: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm.
5. National Health and Medical Research Council, Australian Research Council, the Australian Vice-Chancellors' Committee. National Statement on Ethical Conduct in Human Research (Updated March 2014). Canberra.
6. National Health and Medical Research Council, Australian Research Council, Universities Australia. Australian Code for the Responsible Conduct of Research. Canberra: NMHRC. Available from: https://www.nhmrc.gov.au/guidelines/publications/r39.
7. Australian Government: ComLaw. Privacy Act. Australian Government. Available from: http://www.comlaw.gov.au/Details/C2014C00076.
8. Office of Australian Information Commissioner. Privacy fact sheet 17: Australian Privacy Principles. Australian Government. Available from: http://www.oaic.gov.au/privacy/privacy-resources/privacy-fact-sheets/other/privacy-fact-sheet-17-australian-privacy-principles.
9. Department AGAGs. Information security management guidelines: Australian Government security classification system. Canberra: Attorney-General's Department. Available from: http://www.protectivesecurity.gov.au/informationsecurity/Documents/Australian%20Government%20classification%20system.pdf.
10. Tann J, Flemons P. Our secrets are not your secrets: Proposed national policy and sensitive data report. Atlas of Living Australia. Available from: http://www.ala.org.au/wp-content/uploads/2010/07/ALA-sensitive-data-report-and-proposed-policy-v1.1.pdf.
11. Chapman AD, Grafton O. Guide to Best Practices for Generalising Sensitive Species Occurrence Data. Copenhagen: Global Biodiversity Information Facility. Available from: http://www.gbif.org/resources/2760.
12. Commission AGALR. Australian Privacy Law and Practice (ALRC Report 108). ALRC. Available from: http://www.alrc.gov.au/publications/6.%20The%20Privacy%20Act%3A%20Some%20Important%20Definitions/sensitive-information.
13. Office of Australian Information Commissioner. Information policy agency resource 1: De-identification of data and information. Available from: http://www.oaic.gov.au/.
14. Office of Australian Information Commissioner. Australian Privacy Principles guidelines: Privacy Act 1988. Canberra. Available from: http://www.oaic.gov.au/.
15. Iversen A, Liddell K, Fear N, Hotopf M, Wessely S. Consent, confidentiality, and the Data Protection Act. BMJ. 2006;332(7534):165-9.
16. McGuire AL, Oliver JM, Slashinski MJ, Graves JL, Wang T, Kelly PA, et al. To share or not to share: a randomized trial of consent for data sharing in genome research. Genetics in medicine : official journal of the American College of Medical Genetics. 2011;13(11):948-55.
17. Inter-university Consortium for Political and Social Research. Recommended Informed Consent Language for Data Sharing. ICPSR. Available from: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confientiality/conf-language.html.
18. Inter-university Consortium for Political and Social Research. Guide to Social Science Data Preparation

and Archiving: Best Practice Throughout the Data Life Cycle. Ann Arbor, MI: ICPSR. Available from: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf.

19. University of Western Australia. Guidelines for Drafting a Participant Information Form (PIF) and a Participant Consent Form (PCF). Available from: http://www.research.uwa.edu.au/staff/human-research/human-ethics.

20. UK Data Archive. Example consent form. UK Data Archive. Available from: http://www.data-archive.ac.uk/media/112638/ukdamodelconsent.pdf.

This work is licensed under a Creative Commons Attribution 3.0 Australia License August 2011