ARDC Institutional Underpinnings Element: Sensitive Research Data

30/03/2022

CONTENTS

EXECUTIVE SUMMARY

Institutional Underpinnings is part of the ARDC's National Data Assets Initiative. In this program, 25 Australian universities are collaboratively developing a national Institutional Research Data Management (RDM) Framework. This Framework is intended to inform institutions' design of policy, procedures, infrastructure and services, and improve coordination of RDM within and between institutions. This output describes the initial findings of the research data management Sensitive Research Data Element of the Framework, providing institutions with guidance to ensure compliance with ethical and regulatory requirements, and manage risks associated with sensitive data such as risk to individuals, risk to environments, risk to corporate interests, and legal and reputational risk to the institution. The output recommends the adoption of a "risk and consequence" sensitive data classification scheme aligned to infrastructure platforms with appropriate protections for the different data classification levels. The

output provides an initial step towards consistency of data classification between institutions. The requirement for ethics consideration is also touched on, as is the desirability of sector alignment for sensitive data classifications. Recommendations for institutions and Calls to action are highlighted throughout the Element. Calls to action specifically identify the need for future collective action from institutions and the community. This initial research data management Framework Sensitive Research Data Element will be further developed through additional institutional consultation and will be complemented by activities to validate and test the outputs described within.

DESCRIPTION OF THE ELEMENT COMPONENT

"Sensitive data" can be difficult to define, making it hard to have meaningful discussions about how to manage sensitive research data. We can identify classes of data that are definitely sensitive - for instance, individual patient records, information about the nesting location of an endangered species of bird, or test results that reveal the most effective way to produce a microorganism that could be used as a bioweapon. However, it is harder to agree upon the properties that cause all of these types of data to be grouped into the category "sensitive". In order to sensibly address this area, it is helpful to consider our aims when discussing Research Data Management (RDM) for sensitive data.

For the purposes of institutional RDM, all of the types of sensitive data discussed above are united by the need for special protections and controls to be put in place around their management, including the collection, storage, access and sharing of those data. Institutions are motivated to address the sensitivity of data in order to ensure compliance with ethical and regulatory requirements and manage risk. Data can pose many kinds of risk if mishandled or exposed (including risk to individuals, risk to environments, risk to corporate interests, risk to national security, and legal and reputational risk to the institution itself).

Depending on the kind and severity of the risk posed, different levels of protection and control are appropriate. For this reason, it can be useful to think of sensitivity as a dimensional quality of a dataset rather than a category. There are costs associated with placing protections on data, both financial (more secure infrastructure can be more expensive to purchase and support) and in terms of ease of access and use, so it is not sensible to apply the same protections to all data. More sensitive data require very high levels of protection that are not required for less sensitive data. The degree of sensitivity reflects the severity of the risk posed by the mishandling or exposure of those data.

Viewing sensitivity as a dimensional quality of research data is also helpful when sharing sensitive data for reuse. By its nature, the dissemination of sensitive data must be controlled. But at the same time, the

Australian Code for the Responsible Conduct of Research[1], NHMRC Statement[2], publishers, and funders all encourage data sharing. The level of risk posed by the exposure of a particular dataset can be weighed against the benefits of its reuse when determining whether it should be shared and under what conditions.

The classification of data sensitivity has therefore been identified as a vital component of institutional RDM. Data sensitivity classification schemes set out different levels of data sensitivity, defined according to some set of characteristics of the data. The sensitivity of data according to the classification can be recorded in its metadata. Different protections and handling procedures can be defined for each level.

A long-term goal is for universities to achieve a degree of consistency in classification levels, and lay out the protections required and protections currently in practice at those classification levels. A degree of consistency could simplify decision making around RDM infrastructure and platforms, simplify the use of infrastructure for collaboration, and aid in the movement, sharing and transfer of research data between institutions. Recognition of specific protection levels could further advance trust in the public research sector generally.

There may be two types of challenges that need acknowledging in the collaborative development of sensitive data classifications and protections: the sharing of guidelines for institutions and researchers that are more closely related to the management of data, and the discussion of appropriate technical protections and measures. The latter, technical protections and measures, quickly evolve due to the changing IT, cloud and vendor landscape and the increasing development of cybersecurity.

CALLS TO ACTION

A longer-term goal is to work towards consistency in the recognised protections required and protections currently in practice at certain recognised classification levels. An initial step towards this goal is found in the resources section (see linked document - Classification Crosswalk), an analysis of the typical structure of risk and protection based classification levels across existing examples. For those institutions embarking on classifying sensitive data this could provide some accelerated insights.

Call to action 1: The working group encourages institutions to continue the development of the draft sensitive data classification level crosswalk

[1] "Australian Code for the Responsible Conduct of Research", 2018.
https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018
[2] "National Statement on Ethical Conduct in Human Research", 2018.
https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

Significant benefits could also be achieved by sharing advice on appropriate technical protections and measures, specifically within the context of levels of protection for research data. However, due to the evolving and dynamic nature of the problem, framework documentation may not be an effective mechanism. A preferable mechanism would facilitate communication, dialogue and advice shared amongst institutional infrastructure providers. The ARDC's Australian Sensitive Data Interest Group[3] is one such forum through which these common interests could be explored.

Call to action 2: The working group encourages institutions to come together to discuss cybersecurity protections for Sensitive Data

DIFFERENCES IN APPROACH AND NEED

No Classification vs Build/Adopt

Some universities have no formal data sensitivity classification scheme. Instead, they may provide guidance to researchers about the appropriate protections (and therefore platform choices) through other means, such as within RDM planning tools. Alternatively, an institution may simply use their enterprise information management regulations to assess the protections required for research data.

Choosing not to have a data sensitivity classification scheme avoids the effort required to develop, introduce and maintain such a scheme. However, there are benefits to having a classification scheme that should be weighed against this reduced effort. For instance, a classification scheme allows the sensitivity of a dataset to be clearly recorded within its metadata, making it easier to ensure that the data is appropriately managed and protected. A classification scheme also allows universities to provide broad advice about the protections that are required and platforms that are supported for data of a given level of sensitivity. This is especially useful for universities operating at a scale where one-on-one advice for all research projects is not feasible. Further, it will be easier to assess and work towards increasing consistency in the management of sensitive data across universities if they use comparable classification schemes. For these reasons we encourage the development or adoption of a data sensitivity classification scheme (following the principles outlined below). We also recommend that this scheme be specific to research data (although it may be beneficial to align it with an existing enterprise information management classification scheme). Research data is less uniform and has a different risk profile and user requirements than enterprise data. A classification scheme that was developed specifically for enterprise data is unlikely to be fit for purpose when applied to research data.

[3] https://ardc.edu.au/resources/communities-of-practice/

Recommendation 1: Adopt a Classification Scheme specific to Sensitive Data

Required Protections

Universities vary in the specific protections required for data of a given level of sensitivity. These differences are driven by a number of factors, including:

Location: The regulatory environment surrounding the management of sensitive data varies from state to state. For example, some states require that particular forms of sensitive data be stored onshore, while others require that the same data be stored within the state.

Available Platforms: Universities may be reluctant to impose requirements on researchers that they know cannot be met by the current suite of supported platforms. Additionally, depending on the current infrastructure at an institution, a given level of protection may be considerably more costly in terms of both resourcing for the institution and burden of effort or restricted access on the researcher. These differences will affect the balancing of risk mitigation against cost, and can result in different decisions about the appropriateness of particular protections between universities.

Advice of Legal and Cybersecurity Departments: An institution's understanding of both the regulatory environment that constrains the management of sensitive data and the current best practice for protecting those data depends on expert advice and interpretation. Universities have legal departments and cybersecurity teams to provide this advice and interpretation, and the advice of these groups may differ between universities. How an institution responds to this advice can also depend upon its risk appetite.

Recommendation 2: Define protections required for recognised data classification levels

RECOMMENDATIONS AND ADVICE

Developing or Adopting a Classification Scheme

Clear institutional recognition

The adoption of research data classification will be a significant change to institutional policy or procedures (ref. Policy Element). The Culture Change Element recommends that those leading the change should engage early with research leadership (for example the DVC-R, academic division leadership) to ensure that they are fully aware and in support of the objectives. It's recommended to engage leadership after having a level of maturity and understanding of an institutional risk assessment

and risk management. (ref. Culture Change Element) This will ensure that the change is understood to be important across parts of the organisation, as needs for solutions and process change arise.

Academic institutions can have limited capacity for dedicated work to be done without additional staff and funding. Recognition may be required of the appropriate resourcing and potentially funding for research data infrastructure if not already in place.

Classify only data sensitivity

Universities may require that their research data be classified on many dimensions. For instance, it may be important to record information about the quality of data to inform reuse, or the value of the data to inform retention and disposal decisions. However, it is important to classify these different dimensions separately, as combining too many types of information into a given classification scheme results in classifications that are both more complex and less useful. We recommend that data sensitivity classification schemes reflect only the sensitivity of data, and that other, orthogonal aspects of the data be classified separately.

Classify based on risk/consequences

A potential pitfall is to attempt to define classification levels in terms of the types of information contained within the data (e.g. “data containing personally identifying information”). The problem with this type of classification scheme is that data can be sensitive in many different ways, and the sensitivity of a particular type of information may vary depending on contextual factors and over time. This can lead to very complex definitions of the classification levels that need continuous updating. We therefore recommend that the levels instead be defined by the severity of risk/consequences of mishandling or exposure of the data.

Recommendation 3: Sensitive Data Classification levels should be defined by the severity of risk/consequences of mishandling or exposure of the data

A small but comprehensive set of levels

To ensure broad engagement and applicability the following features are recommended:

- Use an ordinal structure, from the lowest to highest risk/consequences/protections
- Keep it simple, with relatively few levels (many examples have between 3 and 5) - this will make the communications and underlying infrastructure planning as simple as possible. There is a tradeoff between granularity and simplicity - a very complex scheme will allow for fine-tuning of requirements, but if researchers and infrastructure providers find it to be too complex then uptake may be affected.

- Consider how all levels of sensitivity (from open to extreme risk) will be approached. Ideally the full range of levels from open to extreme protection should be covered within the scheme, so that all data can be given a classification. However, many universities are unable to meet the protection requirements of the most high-risk data. Some universities include an "extreme" classification for these data and acknowledge that no local systems are available to provide appropriate protections; others do not include this level, and instead treat data that cannot be managed locally as an edge-case.
- Avoid labelling or defining the levels using language that matches or clashes exactly with other classifications to reduce potential confusion. Universities adopting a scheme from another institution need to adjust the wording to avoid clashes that could occur when the scheme is transposed to their local context.

To ensure the scheme is comprehensive and useful, we recommend aligning the classification levels to research and business requirements and assessing these against existing research services and infrastructure.

Mapping to the institution's information classification scheme

Universities have found it useful to have a simple explanation of the relationship between any overarching institutional information governance, policy or procedure (and potentially with important external information governance such as state government classification schemes). This could outline any agreed relationship by design, relationships between responsibilities, levels, information types, activities or risks outlined in each. This may be most useful for engaging internal stakeholders such as central records, information management or governance groups, IT or cybersecurity.

Supporting Use of the Classification Scheme

Work with/provide guidance to researchers to classify risk

Classifying the sensitivity of data requires the involvement of the researchers who understand that data. However, researchers are likely to need assistance to properly interpret the classification scheme and correctly classify their data.

Both under- and over-classification of data sensitivity is potentially detrimental. Under-classified data will not be given the appropriate protections, increasing the risk of exposure. Over-classification can lead to data being locked down to a point that access and reuse are obstructed. Researchers (and particularly higher degree research students) may be over-cautious in classifying their data early in a project, when they do not yet have a fully-developed understanding of that data. Less-sensitive derivatives of a dataset

may also be developed over the course of a project. It is therefore important that there is sufficient flexibility in the system for classifications to be revised if the researcher later deems it to be appropriate.

Universities may provide assistance with classification to researchers in different ways. Where there is sufficient resourcing, it may be possible to provide one-on-one assistance to the researcher. If the institution operates at a scale where this is not possible, assistance may come in the form of written guidance or interactive classification tools such as surveys or decision trees.

Because data can be sensitive in so many ways, it is not possible to determine that a dataset is not sensitive based on a checklist approach. However, a checklist (or more sophisticated tool) may help researchers to understand the aspects of the data that could lead to increased sensitivity. Additionally, there are some properties that almost guarantee a dataset has a high level of sensitivity that are worth identifying depending on an institution's research profile - for instance, data that contains health information about identifiable individuals. Identifying these properties will help to ensure that the majority of sensitive datasets are correctly classified - but it is important to be clear that such a list cannot be comprehensive, and that the researcher should consider other potential risks and consequences of exposure of their data.

Identify the required protections for each sensitivity level and appropriate platforms to provide these protections

Once the sensitivity levels are identified, universities must determine what the minimum protections are that should be applied to those levels. These protections should ensure that regulatory requirements are met, and should be proportionate to the risk present in the data. Unnecessarily high protections are likely to be both more expensive to run and harder to access, impeding research work.

There should be clear guidance that points researchers to the appropriate institution-supplied (or approved) platforms that meet the protection requirements for each sensitivity level. Researchers cannot be expected to be cybersecurity experts, and may not be able to assess the appropriateness of a given platform against the requirements without expert assistance. However, we do recommend that the protections required at each sensitivity level are clearly stated. This will help more technically-able researchers to contribute to the design of appropriate solutions for edge-case projects, and also may assist with compliance by explaining to researchers why they are restricted to certain platforms for certain types of data.

Recommendation 4: Provide guidance for researchers on the protections afforded through endorsed Sensitive Data infrastructure solutions

Put procedures in place for managing edge-case requests

It is not feasible for universities to provide platforms that cover all use-cases for all levels of data sensitivity. It is helpful to have procedures in place for assessing alternative solutions for these edge-cases, and to understand whose responsibility it is to approve these solutions.

If the sensitivity of the data has been classified (and any additional requirements outlined) then an expert team such as cybersecurity can assess new platforms against the required protections. However, this task involves time and effort, and if the platform is to be used for an extended period of time this assessment may need to be repeated to check for changes that would affect the appropriateness of the platform (for instance, commercial providers may move the location of their data storage from Australia to offshore). For this reason there needs to be a good case for why existing solutions are not sufficient.

In some cases, data may require specific protections which cannot be provided in a cost-effective way by the institution. In this case, the cost of providing the appropriate platform may need to be recovered from the research project. Alternatively, project partners may be able to provide the necessary platform: for instance, if the Department of Defence places certain requirements on the storage of their data, then it may be more appropriate for researchers to access that data from within the Department's own environment.

Assist researchers in securing the end-points

A significant challenge in risk management related to sensitive data (and all cybersecurity) is "securing the end-points". Many data breaches occur as a result of human error or preventable activity, for example theft of laptops or other devices offsite. For this reason, a more comprehensive risk mitigation strategy should be complemented by training and upskilling, guiding researcher practice and instilling good practices in research students. It may be helpful to reach out to IT teams who manage desktop environments for general advice. Some institutions already provide advice for how to better secure personal devices not managed by the institution. If not, consider the potential opportunity for existing IT support to develop and maintain more research focused guidance.

Provide guidance for ethics committees

Not all projects dealing with sensitive data go through ethics committees. However, for those that do, the ethics committee will be concerned that the data are being handled in a way that minimises risk. A clear data sensitivity classification scheme will assist ethics committees to assess whether the planned RDM protocols are appropriate for the level of sensitivity.

For edge cases (data with unusual requirements, or platforms that are not within the suite provided or approved by the institution) it may be harder for ethics committees to assess whether the planned RDM solutions will be appropriate. For this reason, ethics committees should have access to expert guidance -

for instance, by including a RDM expert on the committee, or by giving the committee the option to request an assessment by institutional IT or cybersecurity.

Recommendation 5: Provide guidance for ethics committees on appropriate RDM approaches for sensitive data

Building Towards Alignment in the Sector

Alignment of classification schemes

Aligning classification schemes will make it easier for universities to share and collaborate on sensitive datasets. If universities follow the design guidelines given above, then it should be possible to find the equivalent levels across different schemes. We include as an attachment a draft "crosswalk" between a number of different classification schemes that demonstrates this.

One approach to building towards alignment is to adopt an existing classification scheme from another organisation. When taking this approach, it may be necessary to make adjustments to fit the local context - for instance, to change language to bring it into line with other institutional guidance.

Alignment of protections

What is less straightforward is bringing alignment to the protections that are required for different levels of sensitive data. In the "Approach and Needs" section above we list a number of reasons why universities may differ in the particular protections that they require. It is also difficult to come to agreement between universities on this topic because cybersecurity is a fast-moving area, and best practice will change over time. It is therefore a recommendation of the working group that a community of practice be formed to discuss what protections different universities (and other research organisations dealing with similar data) currently require for sensitive data, changes in best practice, and the extent to which a more aligned "standard" can be developed together.

SETTING EXPECTATIONS

No matter how carefully-designed and sophisticated the classification scheme and risk-assessment tools provided by the institution, the researcher will always share the responsibility for making sure that the sensitivity of their data is well-understood and managed appropriately. The institution is responsible for supporting the researcher in mitigating the risks associated with their data, but cannot be expected to be able to assess and manage that risk without the expert input of the researcher.

Like other aspects of RDM, the management of sensitive data requires ongoing investment. Universities need to monitor changes to regulatory requirements and stay on top of new cybersecurity threats, changes to best practice in the management of sensitive data, and the implications of updates to agreements with commercial providers. Platforms for managing sensitive data will also require ongoing support.

It is not reasonable to expect that the institution will be able to sustainably support the platforms required for all sensitive data use cases as standard offerings. Some data may require unusually high levels of protection that will be too expensive for most universities to implement and maintain. Instead, we recommend identifying and providing for use cases that represent the bulk of research taking place at the institution (for example 80% of cases) and expect to manage the edge-cases separately. This will enable a focus on meeting a smaller number of requirements in a sustainable manner.

It is important to be clear about which requirements are being met by the institution's classification scheme and standard platforms for sensitive RDM. The classification scheme may address the institution's baseline responsibilities, but is not guaranteed to cover additional requirements that may apply to specific datasets (for instance, specific access requirements placed on a dataset by an external provider). Researchers should be given sufficient information about the protections provided by the institution's standard offerings to be able to determine if they require special arrangements to be made to meet any additional requirements.

APPLIED ADVICE

Template Project Plan for an institutional classification scheme

By reframing some of the advice above, the following demonstrates an example of high-level activities that may form an indicative project structure. This structure is not definitive, nor required, but may aid in designing your own structure, or aid in communicating the level of activity and commitment required. Depending on institutional context, this may strengthen an institution's business case for resources.

- Pre-development phase
  - *Typical objectives: to build the case for the change and investment*
  - *Likely activities to be undertaken can include:*
    - institutional risk assessment, general assessment of risk to researchers
    - develop and consult on objectives
    - engage leadership, existing governance, obtain recognition
- Development phase
  - *Typical objectives: broader engagement and buy-in, and to manage expectations with a phased approach, as the extent of implementation may be unknown to begin with*
  - *Likely activities to be undertaken can include:*
    - co-design with researchers & development of levels

    - develop a guided tool, guidance (for researchers, to ethics), governance models
    - formal test of the classification levels and an guidance evaluating usability, inclusivity, and alignment with objectives
    - engage with cybersecurity experts to identify the required protections for each classification level
    - assessment of existing research services & infrastructure
    - recommendations for change (inc. governance, procedures and service changes)
- Implementation phase
    - *Activities vary greatly with existing environments, objectives, and development outcomes, but may involve implementing:* service and support changes, infrastructure changes, training, pathways to appropriate platforms, ways to manage edge cases, ethics processes, ongoing governance.

WORKING GROUP ACKNOWLEDGEMENTS

Nichola Burton (Chair)	ARDC
Lyle Winton (Editorial Committee)	ARDC
Gary Allen	Griffith University
Matthew Bellgard	Queensland University of Technology
Kate Carruthers	University of New South Wales
Jac Charlesworth	University of Tasmania
Penny Cross	University of Wollongong
Kim D'costa	University of Melbourne
Stephen Dart	Monash University
Franco Di Dio	Western Sydney University
Cameron Fong	Sydney University
Owen Griffiths	Griffith University
Ryan McConville	Sydney University
Les Mitchell	University of Southern Queensland
Jen Rowland	Macquarie University
Sharron Stapleton	Griffith University
Jake Surman	University of New South Wales
Milica Symul	University of Canberra
Kandy White	Macquarie University

The outputs of this working group were edited for public release by Frankie Stevens, Lyle Winton and Nichola Burton (ARDC)

Suggested citation: Australian Research Data Commons. (2022, April 1). ARDC Institutional Underpinnings Framework draft release. Zenodo. https://doi.org/10.5281/zenodo.6392340

APPENDIX 1: RESOURCE LIBRARY

Classification Crosswalk

The classification crosswalk spreadsheet (attachment) is an initial analysis of alignment of classification levels across a number of schemes, including some government schemes and some universities.