# RDA SDIG Community Meeting Summary: 12/7/23

Attendees:

* Biru Zhou - McGill University
* Joerg GEIGER - Biobank, University of Wuerzburg, Germany
* Rebecca Tayor-Grant - F1000
* Nichola Burton - ARDC
* Aleks Michalewicz - University of Melbourne
* Rob Baxter - DARE UK
* John Marcotte - University of Michigan
* Matt Cannon - Taylor & Francis
* Monica Marin - Research Institute for Quality of Life, Romania
* Saskia Lawson-Tovey - University of Manchester
* Raoul Vernede - SURF
* Owen Appleton - Swiss Institute of Bioinformatics / BioMedIT.ch
* Dessi Kirilova - Qualitative Data Repository (QDR)
* Lars Eklund - SND(Swedish National Data)/NBIS(National Bioinformatics Infrastructure Sweden)
* Tim Dafoe - Government of Ontario Cyber Security Division
* Gabriela Pino - Universidad Nacional, Costa Rica

Based on the Miro activity at the previous meeting, we focused this discussion on the classification of the sensitivity of research data, with the aim of identifying areas of need for the community. Below we summarise the main discussion points and themes, and identify three potential areas for action.

Note: Group agreed not to publish recordings or verbatim notes to allow for free discussion of how specific organisations handle their sensitive data within the meeting – summary should not identify specific individuals/organisations against comments.

**We asked two questions to prompt discussion:**

**How do you use/interact with data sensitivity classifications in your work? (Does it impact the kinds of data you can collect or research, and the types of systems you can use?)**

* In health environments all data must be classified, from research data to administrative/operational data.
* What can be done with data is directly connected to its sensitivity. So for sensitive personal health data, only actions cleared by an ethics panel, in an approved project and supported by signed legal agreements are permitted.

**What are the challenges to applying these sensitive data classifications? PARTICULARLY when sharing data or working across borders?**

* Every institution has its own sensitivity levels, and national and regional law also have them… You will never get organizations to harmonize their classifications.
* Classifications exist at both the governmental and organisational level

**The following themes arose from the discussion:**

**Disconnect of needs between institutions/IT and researchers when classifying sensitivity**

* Institutional and IT perspective often come from a compliance angle, sometimes leads to overclassification of information, e.g., medium into higher sensitivity, over and above rules and regulations, and does not reflect how research really happens
* Disconnect between institutional, IT and researcher in terms of working with sensitive data
* Also see under classification, sometimes needs to be done by data custodian (or however defined), done by researchers, cannot be reclassified by other units

**Just having the classification scheme isn’t enough - how it’s used is important - both in terms of how data is classified and what is done with data at that classification**

* Not particularly costly for researcher to say ‘this data is not very sensitive’ and to work with the data as they like
* But equally, can overclassify data (lock it away)
* Overclassification risks giving tools to people who want to hide information (e.g., government)
* Can be very hard to push for adequate safeguards for medium or middle range classification (“not that sensitive”
* For individual researchers / areas the sensitivity may be considered differently
* sensitive data does not mean that it is locked down forever, need to consider requirements for ethical reuse, secondary use

**Some features of the scheme itself will have effects on how easy it is to usefully classify data**

* Scheme needs to be simple enough to be used to meet compliance needs / be manageable. That means that classification needs to have limited ‘buckets’, but this means that each bucket could cover quite a wide band

**Researchers may require support in using a scheme effectively**

* Possible to say ‘we are giving researchers a tool against which they can classify their data’ (so, requires effective tool) but there is also a question around whether there should be help for users using the tool to ensure that it is being used correctly
* Risk assessments should not be done by one person or one type of person, requires group comprising specialists in contextual knowledge re what data means plus IT security. If you do not have this then you will be lacking in competencies to complete risk assessment with good outcome.

**A major problem is when the scheme itself has flaws**

* Problems when the schema is bad?!
* Some classification tools are multidimensional flattened
* Unhelpful levels - e.g. Internal doesn’t mean safe or unsafe, but rather administrative, potentially
* We deal with research data as well as other data objects (e.g., encryption keys [‘internal’], etc) > need to make sure that schemas are on a singly dimension or if multidimensional then this be acknowledged > otherwise this is where problems can occur
* A group of Aus universities has provided guidance on design of classifications (Section 7: https://doi.org/10.5281/zenodo.6392340)

**Discussion of whether risk is a single dimension or multiple dimensions**

* Combination of sensitivity and harm into one dimension is a serious issue; we tend to treat things in two dimensions (i.e., harm if disclosed / re-identification risk), e.g., aggregate data about criminal behaviour; phone books
* Harm and re-identification should be approached as different dimensions
* Contextual risk: risk categories can be about risk that information has within context of infrastructure provided; not innate risk of the data itself
* Difficult scenarios where sensitivity does not match our expectation of risk

|  |
| --- |
| **Proposals for activities going forward:**   1. Create guidance about how to create a useful/effective schema (potentially building off the work done in the Australian RDM Framework for Institutions: https://doi.org/10.5281/zenodo.6392340) 2. Analyse differences between schemas that are used in different contexts 3. Provide guidance on how to help people assess sensitivity (everyone will be using a different schema - but approaches to assessing data against that schema could be shared, e.g. decision tools) |