RDA Interest Group Charter

Name of Proposed Interest Group: Sensitive Data Interest Group
RDA site: https://www.rd-alliance.org/groups/sensitive-data-interest-group

Comments on charter and the IG response:
https://docs.google.com/document/d/1Wq-66xfjrLnUhDn5wRZgce8cSRWtsqPpGlYBv6uNLnM/edit

1. Introduction

Sensitive Data: *A working definition of sensitive data is:* Data is information, observations, measurements or some other form of documentation about a place, person, event, animal or phenomenon. It may be digital or non-digital. *Sensitive data is Information that may be regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations. Sensitive personal data* as a specific subset of sensitive data *include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an* identifiable (or potentially identifiable) *individual.* All sensitive data are *data that could potentially cause harm through their disclosure. For local and government authorities, sensitive data is related to security (political, diplomatic, military data, biohazard concerns, etc.), environmental risks (nuclear or other sensitive installations, for example) or environmental preservation (habitats, protected fauna or flora, in particular). Sensitive data may also be subject to* stipulations of customary or communal law of First Nations Peoples, and may be subject to the guidance of communities, traditional leaders, and CARE authority. *The sensitive data of a private body concerns particular strategic elements or elements likely to jeopardise its competitiveness.*
Adapted from: David et al., 2020, "Templates for FAIRness evaluation criteria - RDA-SHARC IG"
https://zenodo.org/record/3922069#.YCJU7ehKg2w

A range of disciplines collect data which are potentially sensitive, presenting serious barriers to reuse and reproducibility. There are a number of barriers which need to be overcome before sensitive data can be utilised safely and to its full potential. One major challenge is that not all sensitive data is alike, with significant disciplinary variation in how sensitive data is defined, linked, managed, stored, and reused. Additionally, common approaches to working with, sharing and managing data are not always appropriate for sensitive data. For example, sensitive data exposes the different perspectives underlying the FAIR and CARE principles. Further, sensitive data requires careful stewarding such that it can be disseminated in an ethically and culturally appropriate way. Nonetheless, sensitive data has significant potential to be utilised in the conduct of novel and impactful work. Therefore, it is essential that a set of community standards and best practices be developed for sensitive data usage and management.

Issues the IG will address
In addition to issues identified by the RDA community as this IG develops, we envisage this IG will address the following issues:

1. Data carries with it different levels of sensitivity depending on its context (e.g., research discipline, who the data is about, what the data is being used for). However, it is not always clear how we should assess data for sensitivities in different contexts. A resource is needed for those working with data to allow them to make informed decisions about data sensitivity and, consequently, data governance, management, and usage.
2. Sensitive data is often de-identified. However, re-identification can be possible and can cause serious harm. Resources are needed on mechanisms of reidentification and the different risks for different types of sensitive data. Such resources could also consider how risk of re-identification may be minimised through the use of "trusted research environments" or "secure research environments".
3. Data that has been labeled sensitive is often not shared beyond the team that collected/created this data. This is a challenge for reproducible research, and means that data collection is sometimes duplicated. More ethically and culturally safe sharing of sensitive data may also enhance the robustness of research design and development. Resources are needed which provide information for those working with sensitive data with information about how that data can be shared and reused in a safe and ethical manner. For instance, a better understanding of the kinds of "secure research environments" or "trusted research environments" used internationally may inform the implementation of sensitive data sharing practices.
4. At times there is a duality between sharing and reusing data in general, and for stewarding data in culturally and ethically appropriate ways (i.e., in general it is good to share data, but there are some cases where data sharing may have negative impacts). This duality is exacerbated in the context of sensitive data due to lower rates of data sharing, and increased potential for harm. Guidelines are needed for balancing principles of data sharing and reuse (e.g., FAIR) with ethically and culturally appropriate principles (e.g., CARE) specifically in the context of sensitive data.
5. Consent is a major consideration when sharing any data, especially sensitive data. However, informed consent can be challenging to obtain, especially when reusing data. This is sometimes a barrier to sharing sensitive data. Guidelines are needed that explore consent models, especially post-hoc consent, for governing the primary and secondary use of sensitive data. Such guidelines should consider the different criteria that must be met to obtain a waiver of consent across different regions.

How this IG is aligned with the RDA mission

The RDA Vision: This IG aligns with the RDA vision because it will develop mechanisms for the responsible reuse of sensitive data - a data source that is both extremely valuable but which also carries many ethical and cultural considerations. Sensitive data will play an increasingly significant role in addressing the grand challenges of the 21st century, such as issues of social and environmental justice and equity, as demonstrated by the impacts of the global COVID-19 pandemic. Indeed, the benefits and potential harms of sensitive data is increasingly being discussed in public fora as corporations and private companies leverage such data for profit. As mechanisms for sensitive data reuse become widely available (such as through the work of this IG), new innovation and invention will be fostered through the reuse of sensitive data. This IG has participants from University and non-University sectors, which strongly positions the IG to engage with all the variety of stakeholders.

The RDA Mission: This IG aligns with the RDA mission as it develops guidelines for the technical components of working with sensitive data, and for addressing the social aspects of working with sensitive data including fostering discussion around the cultural and ethical considerations of data governance and reuse. This IG is well positioned to meet these challenges given the diverse backgrounds of the initial members. The connection between the technical aspects of working with sensitive data (such as secure virtual environments) and the ethical and cultural aspects (such as consent, disciplinary perspectives and norms, and CARE principles) is a key point of interest for this IG.

How this IG would be a value-added contribution to the RDA community
Sensitive data is ubiquitous. However, its context varies. For this reason, this IG complements the work of a range of existing IGs and WGs, including:

- COVID-19 Data Working Group
- Raising FAIRness in Health Data and Health Research Performing Organisations (HRPOs) Working Group
- Reproducible Health Data Services Working Group
- Biodiversity Data Integration Interest Group
- Education and Training on Handling of Research Data Interest Group
- Ethics and Social Aspects of Data Interest Group
- Health Data Interest Group
- International Indigenous Data Sovereignty Interest Group
- Social Dynamics of Data Interoperability Interest Group
- Social Sciences Interest Group
- Infectious Diseases Community of Practice (forthcoming)

The aims of the Sensitive Data IG is to provide a space to focus explicitly on sensitive data. While the scope is interdisciplinary, this IG focuses on sensitive data types. Our planned activities will complement the above IGs as we address sensitive data in domain specific terms (e.g., sensitive data in the health domains) as well as in general terms (e.g., systems for sharing sensitive data). The Sensitive Data IG already has members from a number of the above IGs, which will aid us in coordinating our activities with these groups. The Sensitive Data co-chairs are collectively members of over 20 RDA groups.

All members of the Sensitive Data IG are also active members of the RDA community. We will draw on this to ensure that our efforts take account of previous work in the RDA, and to ensure that our group remains up-to-date on RDA activities.

2. User scenario(s) or use case(s) the IG wishes to address

We identified the following key reasons for forming this IG. We envisage that additional use cases will be developed through working with the RDA community following endorsement.

1. There is a lack of guidelines for working with sensitive data both within and between disciplines/research areas. One reason for this is because sensitive data varies between contexts (e.g., between disciplines). To develop a cohesive but also targeted set of guidelines, a group is needed which comprises members of a range of disciplines with a shared interest in sensitive data.

2. There is a need for a framework which considers the ethical and cultural aspects of sensitive data, alongside the technical aspects. Individuals may want to share their sensitive data and may have conducted all the necessary ethical/cultural safeguards. However, they may lack an understanding of how this can be achieved with the technical resources available to them, what repository or sharing mechanism can handle such data, and how best to access persistent IDs which allow them to track the use of their data. Conversely, individuals may have the ideal technological solution for sharing without an understanding of the ethical/cultural considerations. A group is needed to facilitate a dialogue between the ethical/cultural and technical aspects of sensitive data sharing, and to produce tangible outputs which progress this discussion.

3. There is a general consensus that sensitive data is highly valuable but that it is not being utilised to its full potential. While there is a range of anecdotal support for this claim, a body of work is needed which explores and documents the state of sensitive data primary and secondary usage, and which examines the underlying causes of sensitive data reuse practices within and between disciplines.

4. There is a recognition that there are a number of stakeholders with respect to sensitive data assets, and that each stakeholder has different requirements, needs, expectations, and terminology (e.g., in the case of health data, government, hospitals, researchers, community members). A group is needed which can synthesise the main expectations of different stakeholders to develop resources of individuals and organisations to use when engaging with, sharing, and accessing sensitive data (i.e., a resource for a shared language between stakeholders).

5. There is a need for adequate and specialised resourcing and infrastructure to manage, work with, and share sensitive data. Different data types require different solutions for management, analysis, and sharing. While a range of solutions are available for these different data types, their suitability for sensitive data is not always clear. Work is required to assess solutions for different sensitive data types specifically.

6. Our era is experiencing the most brutal collapse in biodiversity that the earth has known. Yet biodiversity produces many ecosystem services, and resources. However, species and habitat diversity is undermined by many human activities. The preservation of both fragile and overly coveted species and resources makes the publication of their geolocation sensitive. Other data concerning the characteristics of certain pathogens have also proven to be sensitive.

7. The humanities and social science disciplines likewise require clear guidance regarding collection, use and reuse of sensitive data. This may encompass specific ethical considerations pertaining to data collection (e.g., balancing FAIR v CARE principles), research data collection methods when working with marginalised individuals or communities often on sensitive topics, the joining of disparate datasets, and considerations of how long such data should be retained, and where.

8. For biological and geological specimens there is currently a global consultation on extending

core specimen data with data derived from the specimen or linkable to the specimen. This raises concerns related to Access and Benefit Sharing (ABS) and Digital Sequence Information (DSI) in particular for biological specimens, as it would make potentially sensitive data more findable. There seems to be two ways of dealing with sensitive data: obfuscating it to a level at which it is not sensitive anymore but still useful for certain cases, or limiting access to certain experts. Work is needed to develop our understanding of the effectiveness and implementation practices of these methods, especially the latter.

9. In a world of automation, it is likely that there are also automated services generating potentially sensitive data and for these systems a set of automated rules could be developed based on legislation to make a system aware that the data generated is sensitive and deal with it appropriately. For example, a service extracting label data from a biological specimen image (what species is it representing, where and when it has been collected and by whom), might not know whether the data extracted is sensitive.

10. Over time, there may be changes in both the way that sensitive data is defined (e.g., a new data privacy law changes what is considered to be sensitive for a given jurisdiction), and what data is considered sensitive (e.g., a species is newly classified as threatened, or removed from a threatened species list). These historical categorisations of what constitutes sensitive data need to be documented so that records and repositories can be kept current.
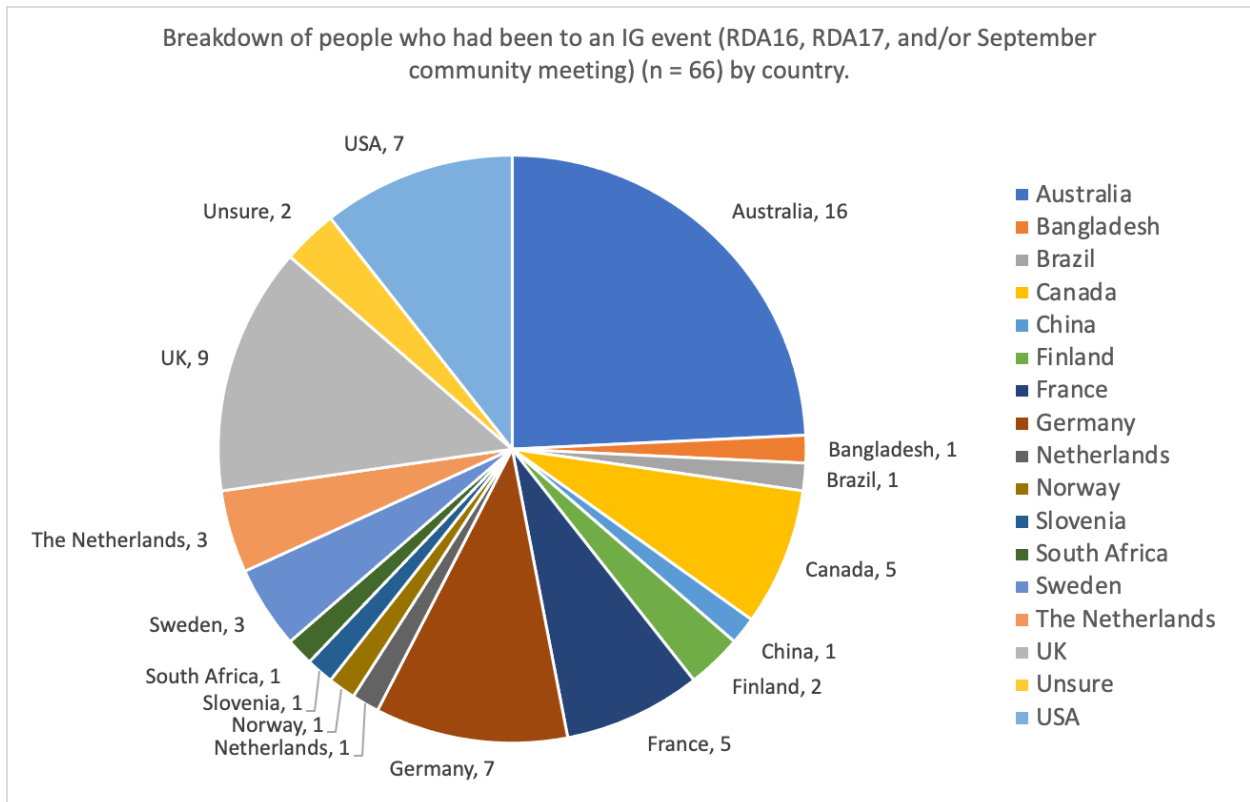
3. Objectives

1. Using the definition presented at the top of this document as a starting point, develop a shared understanding and refined definition of sensitive data.
2. Define various levels of "sensitivity" for data.
3. Data should be as open as possible and as closed as necessary. Within this context, develop an understanding of how sensitivity relates to openness.
4. Identify different consent models.
5. Identify types of sensitive data holdings and resources across various domains.
6. Identify existing data definitions and standards for different types of sensitive data.
7. Identify challenges in collecting, using and sharing sensitive data.
8. Engage with key stakeholders working in the area of sensitive data management/analytics .
9. Identify existing solutions for sensitive data collection, analysis, storage and dissemination.
10. Identify differences in how sensitive data is managed between groups and regions.
11. Document historical categorisations of what constitutes sensitive data and how these change over time

4. Participation

Since the first IG meeting at RDA17, the Interest Group has grown its membership both in terms of disciplinary representation and international participation.
Co-chairs: We currently have 10 co-chairs (7 from Australia, 1 from France, 2 from USA).

Membership: We currently have 69 members signed up to our IG Page (https://www.rd-alliance.org/node/72299/members) from a variety of countries and disciplines. Attendees: Across RDA16 (Sensitive Data BoF), RDA17, and our first community meeting held in September, 2021, we have had 66 people add their name to the attendance sheet from 16+ countries (see figure below) (noting that not all attendees add their name to the attendance sheet, so attendance may be higher). This indicates that the IG is developing good international representation, although Australia still represents a significant proportion of IG members.

Breakdown of people who had been to an IG event (RDA16, RDA17, and/or September community meeting) (n = 66) by country.

Thirteen of those who had added their name to the attendance sheet had been to more than one IG event (e.g., had been to the RDA17 session and the community meeting) and represented 5 different countries (see figure below).

Breakdown of people who had been to more then one IG event (n = 13) by country.
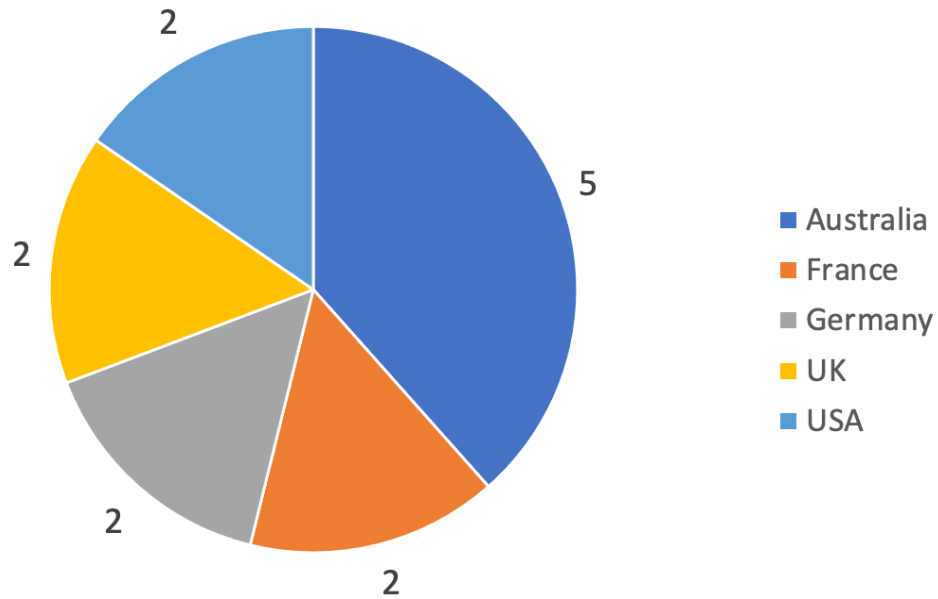
Twenty-nine of those that added their name to the attendance sheet were signed up on our RDA IG page and represent 13 different countries (see figure below).

Breakdown of people who had been to an IG event and are signed up to the IG page (n = 29) by country.
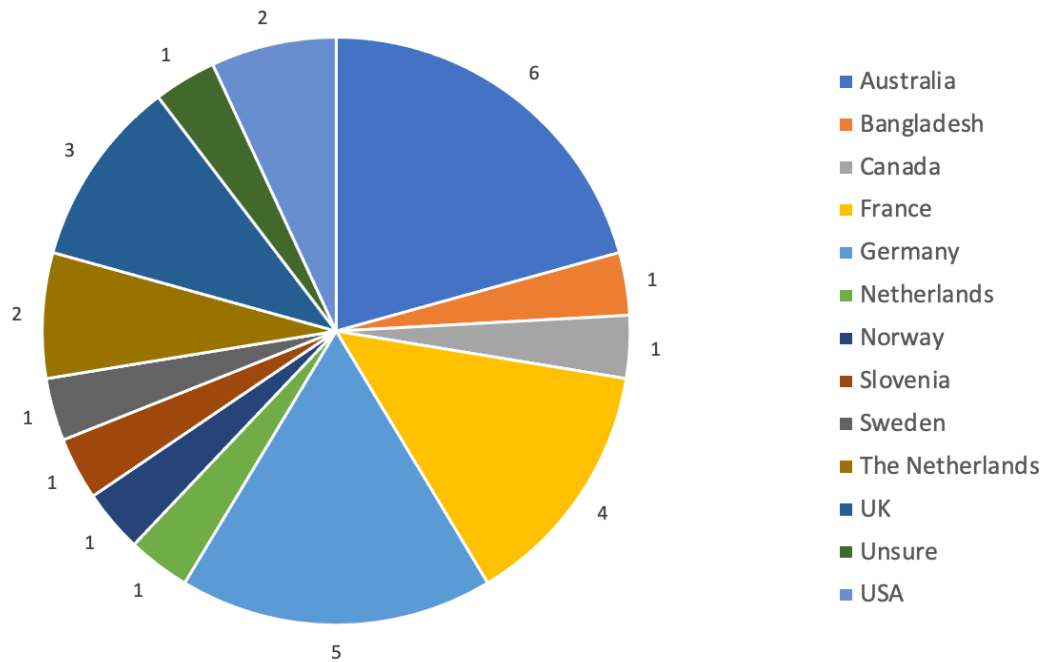
This suggests that the IG is developing a regular community, in addition to the co-chairs, who are interested in this topic.

While participants in this IG are currently mostly from Australia, we have been working to establish this group as part of a global Community of Practice. We are currently developing a strategy to achieve more diversity in international and disciplinary engagement. First, although we can get quite detailed information about attendance and country, we do not have similar information about disciplinary representation. This is something the IG will prioritise getting from members who attend events in the future so that we can track disciplinary representation. Second, the group has seen the recent addition of chairs from Europe and the USA. The Social Science Interest Group, which comprises a broad international membership base and chairs from Norway, USA and Australia, also has formal participation in the Sensitive Data interest group.

The next phase of this strategy will be through specific engagement with RDA groups and other stakeholders covering a range of domains and geographic regions. Specific stakeholders to be approached are still to be determined, but will be drawn from the these target groups which include:

- RDA Interest Groups: Social Science IG (established), International Indigenous Data Sovereignty IG (Initial approach made, pending response), Ethics and Social Aspects of Data IG, RDA-COVID19 WG (and the various sub-groups), Reproducible Health Data Services WG, Epidemiology common standard for surveillance data reporting WG, Domain Repositories IG, Health Data Interest Group, RDA/NISO Privacy Implications of Research Data Sets IG, Virtual Research Environment IG, Social Dynamics of Data Interoperability IG
- Communities outside of RDA: Relevant domain and discipline communities, e.g., the SSHOC and EOSC work programs around sensitive data, US and Canadian networks of Research Data Centres, International and Regional Statistical Agencies (WHO, UNStat, Eurostat, National Statistical Offices)
- Domain specific experts: we will identify domain-specific experts already within our community, and approach external experts, to either join the core group of the IG (as co-chair or as key members of the group) and/or invite them to lead IG initiatives (e.g., WGs) and events (e.g., facilitating workshops run by the IG).

5. Outcomes

1. To identify the key expectations of the community and use these to refine the IG's objectives.

2. List different types of data across disciplines such as health, social sciences, etc, and how different levels of sensitivities apply to those types of data.
3. Identify best practices in sensitive data management across multiple regions, domains and disciplines and how to adapt the best practices.
4. Engage with relevant RDA IGs, WGs and CoPs to identify priorities in the area of sensitive data management.
5. Gather common guidelines and recommendations for working with sensitive data in different disciplines and in different regions, including different consent models.
6. Develop a framework that brings together ethical/cultural aspects with technical concerns of sensitive data.
7. Develop guidelines to aid in assessing the sensitivity of data. The aim of these guidelines is to both ensure that sensitive data are not misclassified as non sensitive (potential for harm) and that data sensitivities are not over estimated (resulting in data not being shared). This also includes examining how sensitive data cna be decoupled from its metadata and the implications of this for sensitive data assessment.
8. Develop guidelines on handling and sharing sensitive data across regions. Sensitive data sharing presents unique challenges to non-sensitive data, and this is especially the case across regions which may have different legislative and legal requirements.
9. Develop frameworks for negotiating different stakeholder expectations when working with sensitive data, including those of communities, public sector, research institutions, and governments.

6. Mechanism

The IG will meet every 3 to 4 weeks via Zoom. Meeting times will be alternated to accommodate as many time zones as possible. Google Docs will be used to develop shared documentation. Email will be used to communicate about meetings and tasks requiring follow-up between meetings. The current co chairs of the IG are already successfully using this system since December 2020 to meet and maintain momentum. More broadly, we have also successfully used this system for Community Meetings - we held our first Community Meeting in September 2021, with the second scheduled for October 2021. Community Meetings will be held slightly less frequently then co-chair meetings; we will aim for every 1 to 2 months depending on group activities.

The IG more broadly will also meet regularly at Plenaries as an opportunity to workshop new ideas with the RDA community and to foster new engagements. The group will also establish an informal communication channel through Slack to allow for ongoing conversation (a slack channel has already been established here "RDA Sensitive Data IG"). The group will also organise webinars and information sessions between Plenaries to share ideas and for group members to stay in touch with the activities of the group. The IG will also use our RDA page to share documents and communicate regularly with the RDA community (the page is already regularly updated with IG activities, see here).

For sharing of resources, the IG has also established a Zenodo community titled "RDA Sensitive Data Interest Group" and a Zotero library called "RDA Sensitive Data ig".

7. Timeline

Initial activities: The group met for the first time as a Birds of a Feather session at RDA 16. Following this, a core group of interested members met to begin drafting the group charter. This group also presented a session and poster for an IG session for RDA 17. The group sent the draft charter for initial TAB review and community consultation in the lead up to RDA 17. Comments on the draft charter have been received and the current group of co-chairs and interested community members are preparing the revised charter for resubmission in October 2021. The IG had our first Community Meeting on September 22, 2021, and has our second Community Meeting planned for October 27, 2021. The IG has had a session and poster session accepted for RDA18. We aim to have the charter approved and the IG endorsed by the end of 2021. Representatives from the IG spoke at an eResearch Australasia Conference (12 October) on Making the Most of the Research Data Alliance (https://conference.eresearch.edu.au/events/making-the-most-of-the-research-data-alliance/). We propose the following work-plan for the first 12 months following endorsement:

1. Formally launch the IG - update our RDA IG site, call for additional co-chairs, share the approved charter with group members, establish a regular meeting time (regular co-chair and community meetings have already been initiated), establish RDA mailing list for the IG.
2. Create resources (format to be determined by the community) aimed at developing a shared understanding of sensitive data through, for example,
    a. surveying definitions and risks of sensitive data in different declipined and regions,
    b. examining varying data sensitivities in areas such as military data, health data, biodiversity data and personal data,
    c. examining sensitive data in different regions (for example the discrepancies which exist in local / national / international contexts,
    d. examining different classifications of sensitive data including dual use of data and applications,

    This work will be undertaken in the first 6-12 months and will be used to help guide other outputs and guidelines (these themes are already being addressed in the IGs session and poster presentation at RDA18).

3. Engage in group consultation to identify the main themes of interest and develop a strategy for establishing working groups/task forces to address these.
4. Engage with stakeholders for feedback on key sensitive data issues and to develop the IGs networks within and outside of RDA.
5. Invite existing RDA IGs identified in section 4 above to provide feedback on, and participate in, working groups/task forces themes.

     a. In relation to points 3 - 5, we aim to focus on the outcomes and objectives listed in section 3 and 5 above.
6. Presentation of webinar/workshop to workshop working group/task force topics and open the working groups/task forces topics for group comment through interactive platforms like Google Docs.
7. Formalise the working groups/task forces, share the goals of the working groups/task forces with the group and RDA more broadly to increase participation, prepare for RDA18 as an opportunity to share progress of the IG and working groups/task forces.
8. Prepare reports and outputs from the working groups/task forces, share reports with the community, present a webinar/workshop to share the outputs with the community.
9. Hold an IG meeting to assess the progress from the preceding 12 months and determine the next steps for working groups/task forces.

8. Potential Group Members

- People interested in leadership

For a list of current and past co-chairs, please see here:
https://www.rd-alliance.org/groups/sensitive-data-interest-group

- People who have joined on the Sensitive Data RDA IG so far

For a list of current members please see here:
https://www.rd-alliance.org/node/72299/members

- People who attended previous RDA16, RDA17 sessions, and the Community Meeting in September 2021, and expressed interest in participating in the IG

For a list of people who have attended past events, please see here:
https://docs.google.com/spreadsheets/d/1S1sEvh1uD4XEqBITVoHaP9z6zle7Wqb10smr_ipQ3V4/edit?usp=sharing