Case Statement: RDA Data Granularity Working Group, Version 1, 5 February 2021

Table of Contents

1. WG Charter / Description

The Data Granularity Task Force of the Data Discovery Paradigms Interest Group (DDPIG) of the Research Data Alliance (RDA) proposes to form an RDA Data Granularity Working Group (WG). This WG would address issues of data granularity in data discovery, access, interoperability, analysis, citation, and more. More efficient and effective reuse of data requires that users can find and access data at various levels of granularity. The WG will explore key questions and collect and share valuable information for how to best support data granularity, providing guidance to help data professionals to determine the best level of granularity for user discovery, access, interoperability and citability. The activities and final recommendations of the Data Granularity WG will build upon and complement existing and ongoing work of several RDA Working and Interest Groups that touch upon the subject of data granularity. The final

deliverable for the WG is a set of collected use cases and a guidance document of data granularity approaches for prioritized use cases, including terminology, methods to evaluate approaches, and a summary of community feedback.

2. Value Proposition

Data infrastructure is generally built around predefined levels of aggregation for datasets and collections, for which conventions vary from one repository to the next. Multiple well-defined levels of granularity can optimize:

- Discovery (by finding the specific data of interest): Enabling discovery at well-defined and multiple levels of granularity can streamline how users navigate. Systems with too low granularity are flooded with entries, especially if these are not grouped by duplicates. This can severely limit the usefulness of the search function. However, when granularity is too high, entries with low granularity are not discoverable by the system, and other ways must be found to enable users to find what they need efficiently. Furthermore, If repositories and/or domains clearly document and organize datasets in a consistent manner, the barrier for new users and interdisciplinary research is reduced.
- Access (by retrieving only the data needed): Ideally, data access can be streamlined to return the specific datasets that the user is requesting, utilizing dataset subsetting and concatenating features which leverage the underlying granularity structure. When granularity is too low, users have to download a great number of datasets and try to fit them together in order to achieve their goals. Conversely, when data granularity is too high, users have to download too much data and then sift through it to get the content that they need. What is too high or too low depends on the domain, discipline and the individual research goal. Furthermore, requirements in the suitability of granularity in a repository may vary between deposit, access, and other services, but repositories are often uncertain as to how to make granularity decisions at these key points, and what information to hand may inform this (e.g., usage statistics).
- Interoperability (by aligning with levels of granularity used in other systems): Interoperability is important for repository networks with shared services or infrastructure, and applications that seek to integrate datasets from multiple repositories. Moreover, even if repositories would like to enable interoperability at a certain level of granularity, the metadata required to do so is not always readily obtained upon deposit.
- Analysis (by readily supporting inputs and outputs of data processing): Choices for data granularity levels should consider analysis outputs such as model runs and data processing results.
- Citability (by providing credit—and further discovery—for the specific data products used): Data citation, which should be in accompaniment of a persistent identifier, needs to be applied at suitable granularity levels to appropriately credit individuals and organizations, precisely represent the data used, and enable interpretable usage

metrics. In terms of credit, a better understanding is needed to determine which situations may or may not lead to having credit propagate between granularity levels. The granularity at which citation is applied impacts the meaning of data usage metrics, and the consistency of these decisions affects comparisons between them. Common approaches are critical to distinguishing relative value of datasets. Citation also relates to dataset versioning practices, explored by many RDA groups.

- Curation and Deposit (via both manual and automated workflows): Curation tools and workflows have influenced data granularity decisions. In some cases, splitting data into smaller datasets is more time consuming (e.g., lack of ability to clone metadata as a starting point from a similar dataset, limited tools to associate related datasets together into collections), and in other situations publishing smaller parts as they are ready is more efficient. Researchers submitting their own datasets may have a greater tendency to take the simplest approach to meet deposit requirements, as opposed to curators who are attempting to facilitate data services that rely on granularity decisions. Quality metrics and processes are hard to define without specifying the granularity level they apply to. What may be appropriate for small scale measurements may be unreasonable for large conglomerates of heterogeneous data sources and vice versa.

Rich information can be gained by working with data at various levels of granularity (e.g., collections, datasets , observations, and more). More efficient and effective reuse of data requires that users—be they humans or machines—can work with data at various levels of granularity. For example, within a data collection composed of a set of files, researchers could explore within each file variables, geospatial layers, or individual observations. To enable these levels of discovery, analysis, and citation, supporting greater granularity can include the creation and support of metadata and persistent identifiers at various levels of granularity. At present, some repositories support data granularity, e.g., the ability to discover, query, and access files within a collection, layers in a complex file, or columns in a table.

The WG will collect and share valuable information for how to best support data granularity, providing guidance to help data professionals to determine the best level of granularity for user discovery, access, interoperability and citability. The WG also will explore key questions regarding:

- optimal levels of granularity for varied usage contexts,
- storage and distribution of different levels of granularity,
- how should different levels of granularity (within the same dataset) be presented to the user,
- relationships amongst levels and with other datasets,
- tracking the granularity of dynamically-generated data, and
- common terminology for granularity concepts.

Given the integral, cross-cutting nature of granularity across a range of issues, a prime value of the WG will be leveraging and building upon existing and ongoing RDA work amongst a range of groups (see below).

The key beneficiaries of the WG activities are as follows:

Beneficiary	**Tangible Impacts**
Data producers	● More efficient data management ● Greater ability to track value and use of subsets of data
Data infrastructure providers (e.g., repositories)	● Enhanced user services ● Clarity on the levels of data granularity on which to operate (for deposit, storage, and access) ● Greater ease in exchange of subsets of data (and metadata) with other systems
Secondary data users	● Faster discovery and access of data of interest ● Expanded research possibilities given new ways to work with granular datasets
The public	● Benefits from faster and more sophisticated research produced

3. Engagement with Existing Work in the Area

Engagement with Adjacent RDA Groups

The activities and final recommendations of the Data Granularity WG will complement the work of several RDA Working and Interest Groups that touch upon the subject of data granularity and present use cases. These RDA groups have been identified by scanning their aims and outputs, of which the relevant WG/IG and outputs are listed in detail in **Appendix A**. The risk of duplicating work is low as this preliminary search through the RDA outputs concluded that while some of the outputs touch the topic of data granularity, none of the existing groups cover it in detail or provide recommendations. Where relevant information is discovered, the WG will leverage that work in its recommendations. In order to ensure communication between the relevant active IGs/WGs, we will connect to the identified groups both by the respective mailing lists, as well as by inviting members of the groups to join this WG calls (incl. Plenary sessions).

Additionally, we will establish liaisons with adjacent RDA groups, especially in cases where activities have a strong complementarity.

4. Work Plan

4.1. Final Recommendations

The final deliverable for the WG is a set of collected **use cases** and a **guidance document** of **data granularity approaches** for prioritized use cases, including terminology, methods to evaluate approaches, and a summary of community feedback.

4.2 Deliverables & Milestones

0-6 Months:

Define use cases: While some use cases were identified during preliminary work as a task team prior to becoming an established working group, these need to be properly documented, expanded and prioritized. A template for recording use cases will be identified as an early step (such as those listed at https://www.rd-alliance.org/group/research-data-collections-wg/wiki/research-data-collections-use-cases). Ideally, this work will be completed by around month 3.

Identify existing constraints within registries, metadata standards and relevant existing RDA WG outputs where these data granularity concepts are applied. We will look to identify relevant metadata fields, commonalities and differences, and issues. Some examples are listed below to illustrate the types of information that will be examined.

- Registries examples:
    - DataCite has relationships like ‘IsPartOf’ that may provide the linkage between a dataset and its parent collection
    - IGSN has concept of hierarchical collections, samples and sub-samples
    - OBIS has a schema denoted as OBIS-ENV-DATA that has archives organized with events, occurrences (of marine organisms) and their related ‘measurements or facts’.
    - RADAM gives DOI for queries on underlying databases, so they can be citable
    - OpenCitations only registers data citations that use a DOI
    - WDC for climate has very coarse granularity. CMIP6 data for IPCC reports has a very fine granularity, which makes it hard to map between the two.
    - General purpose repositories often don’t have the expertise to decide upon an appropriate granularity.
- Metadata standards examples:
    - ISO 19115 includes a field for parentMetadata that can allow for hierarchical relationships of datasets

    - DDI (Data Documentation Initiative) includes a variety of granularity levels, such as variables and variable groups, and hierarchical relationships.
- RDA WG output examples:
    - RDA Research Data Collections metadata schema
    - Data Usage Metrics recommendations

Conduct an initial review to identify aspects of granularity from public resources available on the repositories.

Create and conduct a survey of the community to solicit:

- existing approaches to segmenting datasets and collections in repositories (e.g., what levels of granularity are they operating at in different functions);.
- end user expectations for access and discovery; and
- known problems and issues from WG team members and related RDA groups (via identified liaisons).

6-12 Months

- Summarize and analyze findings from the first 6 months.
- Collect additional data if necessary, and consolidate with the community through targeted communication tools (webinars/gdoc/BoFs etc)
- Draft guidance paper which addresses prioritized problems/issues, survey results and domain considerations.

12-18 Months

- Feedback will be solicited from relevant stakeholders on the draft guidelines (likely tools: webinars, RDA Plenary sessions, google docs, surveys, comments on posts).
- If available, integrate outcomes of early adopters or pilot implementations that emerge from the adoption plan.
- Finalize, publish and promote guidance paper.

4.3 Working Group Operations

The three co-chairs will share the responsibilities for organizing meetings and plenary sessions. Specific task groups will be established to progress on identified deliverables, with leads that will report progress at team meetings. Regular working group meetings will be conducted monthly in two time zones to encourage international participation, using video conferencing and Google drive meeting minutes. A mailing list will be used for asynchronous communications. Other tools may be used for collaborating on and tracking deliverables.

4.4 Community Engagement and Participation

Broader community engagement will be required at multiple stages, with the survey and draft guidelines feedback being the most significant. The survey will be advertised to relevant RDA WGs, the PID Forum (https://www.pidforum.org/), and other research data management mailing lists/communities (e.g., ESIP, AGU ESSI, etc). Survey results and the draft guidelines will be presented to get constructive feedback, with engagement details distributed to the same communities as the survey. Feedback will be categorized, evaluated and considered for integration.

5. Adoption Plan

The Granularity WG will define the key assumptions and concepts for data granularity, which will be translated into guidelines and best practices for capturing data granularity for their widespread adoption across the different stakeholders in the data lifecycle; data producers, data curators, data managers and data users are among the key roles that stand to gain. The WG will engage with these communities at national, disciplinary, and international levels.

The WG will organise dissemination about the activities and findings and gather community feedback regularly during all the phases of the work. To promote transparency and accessibility of work in progress, the group members may utilise a public GitHub repository for WG collaborative documents. The following table details the engagement that will be undertaken towards relevant milestones, including final adoption.

	Milestone	*Engagement*
1	Identify issues in definition of data granularity (domain agnostic and domain specific), based on analysis of existing definitions and frameworks.	As per survey in above Work Plan.
2	Draft plan of guidelines, including identification of use cases, and community consultation on guidelines.	After gathering use cases and reviewing challenges regarding the granularity aspect, the WG will produce a draft best practices document, which will be circulated and validated with the community for feedback.
3	Complete community consultation and finalise definition of guiding	

	principles for data granularity	
4	Draft plan for adoption of guidelines, including identification of adoption examples	Webinars, initiate possible pilot implementations
5	Implement plan for adoption of guidelines, which will continue after the cessation of this WG	The adoption plan will address how to work with different stakeholders, including: ● those that will endorse and promote the guidelines ● those that will provide training on the guidelines ● users of the guidelines ● and will include suggestions on follow-up work

6. Initial Membership

FirstName	LastName	Affiliation	Country	Member Type
Katherine	McNeill	Harvard University	U.S.	WG Co-chair
Reyna	Jenkyns	Ocean Networks Canada	Canada	WG Co-chair
Brigitte	Mathiak	GESIS - Leibniz institute for the Social Sciences	Germany	WG Co-chair
Esther	Plomp	Delft University of Technology	The Netherlands	
Fotis	Psomopoulos	Institute of Applied Biosciences, Centre for Research and Technology Hellas	Greece	DDP IG co-chair
Maggie	Hellström	Lund University	Sweden	

		and Integrated Carbon Observation System (ICOS)		
SiriJodha	Khalsa	Univ. of Colorado	U.S.A.	DDP IG co-chair
Graham	Smith	Springer Nature		
Luc	Decker	Institut de Recherche pour le Développement	France	
Marie-Lise	Dubernet	Paris Observatory	France	
Thomas	Jouneau	Université de Lorraine	France	

Appendix A: Engagement with Existing Working and Interest Groups

RDA W/IG	Status	Description W/IG	Output (and relevant pages)	Output Summary
Data Citation WG	Maintenance (started in 2015)	The Data Citation WG aims to establish best practises for efficiently identifying and citing arbitrary subsets of (potentially highly dynamic) data	Scalable Dynamic-data Citation Methodology (2015)	Developed a simple, scalable mechanism that allows the precise, machine-actionable identification of sub selections of data, irrespective of any subsequent addition, deletion or modification. Data should be versioned and assigned PIDs to timestamped queries/data.
Data Discovery Paradigms IG	Established (since 2016)	The DDPIG aims to improve data discovery.	Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories (2017) - p. 9, 13-15	Granularity was discussed as an important aspect of required metadata. Recommendation 3 addresses some aspects of granularity during users' judgement of dataset fit for their specific use case.
Data Fabric IG	Established (since 2014)	The goal of DFIG is to identify common components and define their characteristics and services that can be used across boundaries in such a way that they can be combined.	Not applicable.	Some relevance, but then related to digital objects being assigned PIDs that 1) resolve to machine-actionable resources (e.g. landing page) and 2) are able to be associated with specific kernel information that encodes e.g object type in machine-interpretable ways.
Data in Context IG	Established	The DiCIG aims to set up contextual profiles, with documentation of the evolution of the data asset behind each element.	Not applicable.	The contextual profiles seem to relate closely to the concept of versioning (which I expect can relate to granularity, how different subsets of data were gathered and joined or separated over time). Focus seems to be the idea of contextual

				metadata.
Data Type Registries WG & #2	Paused	DTR WG activity is currently paused while the topic of data type registries is under consideration by ISO as a potential standard. The next step in DTR work is governance.Detailed and precise data typing is a key consideration in data sharing and reuse and that a federated registry system for such types is highly desirable and needs to accommodate each community's own requirements.	Data Type Model and Registry - Data Type Registries (DTR) WG Recommendations	Not applicable.
Data Usage Metrics WG	Established - Wrapping up	The DUM WG aims to harness community buy-in of data usage metrics and drive widespread adoption.	Code of practice for research data usage metrics release 1 - p. 3, 24	Granularity report attribute via API for reporting periods.
Data Versioning WG	Established - Wrapping up	The DV WG aims to establish standards for data versioning so that specific parts of datasets can be cited.	Principles and best practices in data versioning for all data sets big and small (2020) - p. 2, 3, 10, 11, 13 Compilation of Data Versioning Use cases from the RDA Data Versioning Working Group (2020)	This output focuses on data versioning, touches upon the subject of data granularity and provides use cases that can be helpful to the WG. 39 use cases: A) Web sources (use cases 1-10); B) RDA Sources (use cases 11-12); and C) Data Repositories (use cases 13-39).
FAIR Data Maturity Model WG	Maintenance	The FAIRDMM WG aims to establish common set of core assessment criteria for FAIRness and a generic and expandable self-assessment model for measuring the maturity level of a dataset	FAIR Data Maturity Model: specification and guidelines	This output does not mention granularity but the work of the DG WG may contribute to improving the FAIRness of datasets.
Metadata IG	Established	The Metadata IG will concern itself with all aspects of metadata for research data	No RDA outputs	Not applicable.
Metadata Standards for	Established - Wrapping up	The MSAPDCS WG will address the incomplete	RDA/TDWG Attribution	Has a list of aggregating metrics for research

attribution of physical and digital collections stewardship		standards for giving attribution for the maintenance, curation, and digitization of collections.	Metadata Working Group: Final Recommendations - p. 2	products.
Physical Samples and Collections in the Research Data Ecosystem IG	Active (since 2017)	Focus on persistent identifiers and metadata for physical samples	No RDA outputs	Not applicable.
PID IG	Established (since ?)	The PID IG will define emerging PID use cases in the domain of data, and whether the research community would benefit from a global open identifiers for persons, data objects, organizations, grants, etc.	No RDA outputs	Not applicable.
PID Kernel Information Profile Management WG / PID Kernel Information WG	Established - Wrapping up	Continuation of the PID Kernel Info group, more focused on governance issues.	Recommendation on PID Kernel Information - p. 8, 10	Unclear if it's relevant. Aims "to advance a...change to middleware infrastructure by injecting a tiny amount of carefully selected metadata into a...PID record". There is some discussion of documenting a data file that is related to/part of a dataset (P 8), and the Kernel itself has some version attributes (p. 10), but nothing significant on granularity.
Preserving Scientific Annotation WG	Not yet endorsed	Not applicable.	Not applicable.	Not applicable.
Raising FAIRness in health data and health research performing organisations (HRPOs) WG	Established - Getting started	This is a new WG focusing on guidelines to apply FAIR principles for health and clinical research data. The granularity of the data to be shared is an important consideration.	Interesting outputs in the future.	Not applicable.
RDA/FORCE11 Software Source Code Identification WG	Established - Wrapping up (started in 2019)	The RDA/FORCE11 SSCI WG aims to bring together stakeholders directly involved in software identification	Software Source Code Identification - p. 9-10, 13-16,	Focus on which persistent identifier is suitable for which level of software granularity, with ARK as winner
RDA/WDS Scholarly Link	Maintenance	This WG is the follow up from the: RDA/WDS	Scholix Metadata	Could relate to granularity if we consider the linkages

Exchange (Scholix) WG		Publishing Data Services WG working towards a global information commons.	Schema for Exchange of Scholarly Communication Links	between other products (e.g., articles) and datasets at a more granular level. The outputs of the DG WG could change the specification to look at the more granular level if it would be useful in linking that level to other scholarly outputs.
Reproducible Health Data Services WG	Established - Getting started (started in 2019)	The goal of the working group is to enhance the reuse of health data for research and improve the FAIRness levels of aggregated and curated data sets for secondary use.	Interesting outputs coming up in the future.	Not applicable.
Research Data Collections WG	Maintenance (started in 2015)	The PID Information Types WG has defined a core model and the central interface for accessing object state information and provided a small number of example types, which were consequently registered in the Type Registry WG prototype.	Final report/recommendations of the RDA WG on Research Data Collections (2017)	These recommendations are relevant for situations when “granules”(possibly from completely different sources) need to be aggregated into new unit, as they provide stringent instructions for identifying each collection item (preferably using their PIDs).
Research Data Repository Interoperability WG	Maintenance (started in 2015)	The RDRI IG will establish standards for interoperability between different research data repository platforms.	No RDA outputs yet	There is potential for a future connection (one could develop interoperability at a more granular level, but the focus to date still is at the ‘study’ level).
Research Metadata Schemas WG	Established - Wrapping up (started in 2019)	The RMS WG aims to identify gaps in existing schemas commonly used for research data and to provide guidelines for those communities whose needs are not addressed by existing metadata schema such as schema.org.		There might be potential for a future connection if systems like Schema.org could be further expanded to describe (and thus facilitate discovery) at a more granular level, but the schemas used now cover just the ‘study’ level.
WDS/RDA Assessment of Data Fitness for Use WG	Established - Wrapping up (started in 2017)	The WDS/RDA ADFU WG aims to establish standards for data quality and preferably a corresponding metric.	Checklist for Evaluation of Dataset Fitness for Use - p. 4	A checklist supplementary to the CoreTrustSeal requirements were generated to manually determine the fitness for use of datasets (at least a subset) within a repository. One of the criteria in the checklist is to determine if the granularity of data

				entities in the dataset is appropriate. There is a reference to a google document (see p.2 and 3 of Guidelines in Respect of MetaData Granularity (see goo.gl/komKJz); unfortunately the document is unavailable for viewing.