Guidance on Data Granularity

Report of the RDA Data Granularity WG

Publication Date: September 2024

Authors:

Name	Affiliation	ORCID	CRediT Contributor Roles[1]
RDA Data Granularity Working Group	Research Data Alliance	n/a	Conceptualization, Methodology
Reyna Jenkyns	World Data System–International Technology Office	0000-0001-6975-6816	Investigation, Supervision, Writing – original draft and review & editing
Brigitte Mathiak	GESIS	0000-0003-1793-9615	Formal Analysis, Investigation, Supervision, Writing – original draft and review & editing
Katherine McNeill	Independent	0000-0003-2865-3751	Formal Analysis, Investigation, Supervision, Writing – original draft and review & editing
Graham Smith	Springer Nature	0000-0001-9520-0109	Formal Analysis, Writing – original draft
Guangyuan Sun	Nanyang Technological University–National Institute of Education	0000-0001-7352-2158	Formal Analysis, Writing – original draft and review & editing
Chris Little	Meteorological Office	0000-0002-1442-3712	Writing – review & editing
Beverly Jones	University of Sheffield	orcid.org/0000-0002-3900-9374	Investigation
David Elbert	Johns Hopkins University	0000-0002-2292-180X	Investigation
Rouven Schabinger	Swiss Library Service Platform	0000-0002-0249-7917	Investigation
Maggie Hellström	Lund University	0000-0002-4154-2610	Visualization

[1] CRediT (Contribution Roles Taxonomy) (National Information Standards Organization, 2024)

Title: Guidance on Data Granularity

Abstract: Data infrastructure can be built around predefined levels of granularity for data, for which conventions vary. The appropriate level of granularity can optimize discovery, access, interoperability, analysis, identification, citation, curation, and more. This guidance document developed by the Research Data Alliance (RDA) Data Granularity Working Group (WG), along with its Supporting Output Use Cases, provides guidance on data granularity approaches, issues to consider, and priority use cases for research data infrastructure providers and other key stakeholders.

Contributions to the United Nations Sustainable Development Goals (SDG): Directly covered: 9 (Industry, Innovation, and Infrastructure), 12 (Responsible Consumption and Production), 17 (Partnership for the Goals). Indirectly covered: all others.

Language: English

Version: 1.0

Keywords: access, aggregation, citation, identification, data granularity, digital artifact, disaggregation, discovery, FAIR principles, interoperability, repositories, use cases

Executive Summary

Data infrastructure can be built around predefined levels of granularity for data, for which conventions vary. The appropriate level of data granularity can optimize discovery, access, interoperability, analysis, identification, citation, curation, and more. Many data infrastructures can benefit from further guidance on these use cases and how to optimize their systems and processes for granularity. Furthermore, terminology related to data granularity can vary widely, leading to confusion or difficulty in interoperability.

The Research Data Alliance (RDA) Data Granularity Working Group (WG) was created to address these needs. Building on prior RDA efforts, the WG explored how to best support data granularity across infrastructure and services, through creating use cases; surveying the data repository community; and conducting an environmental scan of existing constraints within registries, metadata standards, and relevant existing RDA WG outputs. Investigations yielded the following core recommendations, elaborated in the Recommendations section:

- **Recommendation 1, Entity Characteristics:** Entities should possess a key set of characteristics and relationship properties.
- **Recommendation 2, Metadata:** Infrastructure should make use of metadata to support and enable documentation of data granularity.
- **Recommendation 3, Data Policy:** Policies should be put in place to support data granularity.
- **Recommendation 4, Discovery & Access:** Infrastructure providers should enact ways to enhance discovery and access at multiple levels of granularity.
- **Recommendation 5, Identification & Citation:** Infrastructure should put into place mechanisms that support the identification, citation, and tracking of entities at multiple levels of granularity.
- **Recommendation 6, Examination:** Stakeholders should proactively examine and assess how best to support data granularity in their context on a periodic basis.
- **Recommendation 7, Terminology:** Stakeholders should adopt the terminology recommended in this document, in order to facilitate common understanding and interoperability of granularity.

This guidance document, along with its Supporting Output Use Cases, provides recommended terminology, guidance on data granularity approaches, issues to consider, and priority use cases for research data infrastructure providers and other key stakeholders.

Table of Contents

Executive Summary **3**
Table of Contents **4**
Introduction **4**
Background 5
Data Granularity Terminology 6
Methodology **11**
General 11
Use Cases 12
Survey 13
Constraints Evaluation & Environmental Scan 14
Findings **15**
Use Cases 15
Frequency statistics 15
Challenges arising 19
Survey 20
Demographics 20
Entities 22
Policies 24
Constraints Evaluation & Environmental Scan 26
Metadata schemas 26
Endorsed RDA Outputs 28
Overall Trends and Observations 30
Recommendations **31**
Conclusion **35**
Acknowledgements **36**
Appendices **37**
Appendix 1: Documents from Related RDA Groups 37
Appendix 2: Schemas and Registries Reviewed 39
Appendix 3: Use Case Spreadsheet Documentation 40
Appendix 4: Repository Survey Invitation 41
Appendix 5: Repository Survey Instrument 42
References **57**

Introduction

Data infrastructures generally are built around predefined levels of granularity (or aggregation) for data, for which conventions vary. More robust information can be gained by working with data at various levels and dimensions of granularity; for example, more efficient and effective reuse of data requires that users can find and access data at various levels of granularity. The appropriate level of granularity can optimize:

- discovery, because it helps users more efficiently find the data required;
- access, because it delivers the needed data without having to involve further subsetting at the user's end;
- interoperability, because it enables exchange between infrastructures, including combination of data from different sources;
- identification, because it can designate a specific data component;
- citation, because credit can be attributed to the producer and publisher easily according to the right level of data; and
- curation, because it enables more tailored management of specific sets of data.

Many data infrastructures can benefit from further guidance on these use cases and how to optimize their systems and processes for granularity. Furthermore, terminology for key concepts related to data granularity can vary widely, leading to confusion or making it more difficult for stakeholders to interoperate. The Data Granularity WG was created to address these needs.

The key intended beneficiaries of the WG's activities: data producers, data infrastructure providers (e.g., repositories), and data consumers. The final deliverable for the WG is a set of collected use cases (see Supporting Output) and this guidance document of data granularity approaches for prioritized use cases, including terminology, methods to consider approaches, and a summary of community input and feedback. The purpose of this document is to provide guidance for key stakeholders (e.g., data managers, infrastructure providers, data repositories) to help them determine the best levels of granularity to optimize discovery, access, interoperability, identification, and citability, among other tasks.

Background

In 2019, members of the Data Discovery Paradigms Interest Group began to track the significance of data granularity across a variety of aspects of data infrastructure (including not only discovery, but also versioning, identification, metrics, and citation). It established a Data Granularity Task Force, which undertook tasks to identify existing data granularity problems and questions, understand use cases, link with related RDA groups, grapple with terminology, yielding an internal white paper (RDA Data Discovery Paradigms Granularity Task Force, 2019a). Given the complexity of the issue, and its extensiveness far beyond the concept of data discovery, the Task Force initiated the proposal for a formal and independent RDA Data Granularity Working Group (WG), intended to further this work and provide formal recommendations.

Community members interested in such a working group began meeting in 2020 to begin its definition and launch process. The Working Group's initial Case Statement was developed and shared in February 2021; based on community feedback, the second and final version was published in July 2021 (Research Data Alliance, 2021). The Data Granularity WG was charged with addressing issues of data granularity holistically, across the realms of data discovery, access, interoperability, analysis, identification, citation, and curation and deposit. As articulated by stakeholders in the Case Statement, the value of this work will:

- "foster the use cases where citation and reuse of data require a more precise and flexible level than the dataset";
- "enable better pathways for reuse, citation, credit, and incentivization for FAIR and open research data"; and
- serve as "an important building block towards enhancing data findability and thus crucial on the way to FAIR data."

As the WG completed its tasks (see Methodology), its activities reinforced the value of having common terminology to discuss granularity issues–not only among the broader community (as originally intended), but also amongst members of the WG itself in order to produce the recommendations. The section that follows outlines the group's recommendations on terminology, shared both to clarify usage in this document and for broader community adoption.

Data Granularity Terminology

Contending with terminology is a key issue in data granularity. Stakeholders in the field do not necessarily use consistent terminology, leading to potential confusion or missed opportunities for harmonization and interoperability. The WG puts forth the following definitions for common usage when supporting and discussing granularity in data infrastructures, as a result of their work and review of common terminology in the field for granularity concepts (Albertoni et al., 2024; Australian Bureau of Statistics, 2023; Cambridge University Press, 2024; CODATA, 2017; DataCite Metadata Working Group, 2021; DDI Alliance, 2024; Earth Science Data Systems, 2015, 2015; eurostat, 2024; Gregory et al., 2018; ICPSR, 2024; Keet, 2010; Murphy & Parsons, 2021; RDA Data Discovery Paradigms Granularity Task Force, 2019b, 2019a; RDA Data Granularity Working Group: Use Cases Subgroup, 2024; Stevenson, 2010; United Nations, 2024, 2024; Wu et al., 2019).

This set of terminology is not perfect (as none would be), and alternative vocabulary has been used and debated (even within the WG). Moreover, this paper does not propose a requirement for this terminology to be used. That said, the most important thing to advocate is to use some kind of consistent terminology, to enable common understanding and interoperability amongst stakeholders and infrastructure.

Granularity in general can be defined as "the scale or level of detail in a set of data" (Stevenson, 2010) or "the ability to represent and operate on different levels of detail in data and information" (Keet, 2013). The group identified two main kinds of data granularity:

- *Component granularity*, wherein a set of data can be subdivided or aggregated into smaller or larger components. For example, survey data can be delivered as an entire data file, or just the answers for a specific respondent.
- *Dimension granularity*, wherein units can be measured (and data represented) according to finer or coarser units of measure. For example, time can be recorded at finer (minute, second) or coarser (year) levels of detail. The units are described relative to that level, which may be more coarse-grained or concern fine-grained details.

Either of these types of granularity can be represented in a hierarchical structure, wherein one level is a subcomponent of another (Figure 1) or where a finer level contains data at a greater level of precision (Figure 2). Whilst this paper primarily discusses component granularity, many of the recommendations would help to better support dimension granularity as well. The hierarchy can be constructed through granulation (decomposition of whole into parts) or aggregation (integration of parts into a whole).

Some arrangements of data may not map to a single level of granularity per se; for example, a survey may contain some variables regarding the household level and others reflecting the level of an individual. Such arrangements create new ways of understanding the data that might not fit neatly into a single level of detail in the original data structure.Thus it is important to highlight that data may be grouped and packaged independent of concepts of hierarchical granularity. Additionally, not all datasets or infrastructures in the field will employ all of these concepts and terminology. Lastly, data can be represented at multiple levels of granularity.

Table 1: Granularity Terminology

Proposed set of common, defined concepts to be used when discussing data granularity.[2]

Table 1a: Levels of Component Granularity (aka Entities)

Term	**Description**	**Sub-component of**	**Alternative Term(s)[3]**	**Examples**
Data point	The individual result of data collection, at the smallest unit of data which can be retrieved.	Variable	Data item Datum Granule Observation Measurement	Single number or value
Variable	A quality being measured across units in a dataset. A series of single values	File, Database	Field	Column in spreadsheet Responses

[2] Documentation for the WG Supporting Output (namely the Use Cases) include scope notes and definitions for other relevant concepts, such as stakeholder types.

[3] Readers may see these other terms used, however it is not recommended to use different terms interchangeably.

Term	Description	Sub-component of	Alternative Term(s)[3]	Examples
	measured in the same way.			to a common survey question
Database	A set of data stored on a computer system, often in a way that allows search and retrieval of a set of data based on criteria.	Dataset (sometimes)	Relational database Datastore Repository	
File	Digital object for storing measurements on a computing device, identified by its filename.	Dataset (if it contains more than one file)		Spreadsheet Image
Custom subset	A user-created selection of part of a dataset. May or may not be independently identifiable, discoverable or citable, depending upon the infrastructure.	Dataset		
Dataset	A package/organized set of data deliberately published and represented by a metadata record in a data catalogue. May comprise multiple files (of data, metadata, and possibly related files such as code) and/or in database format. According to DCAT a “collection of data, published or curated by a single agent, and available for access or download in one or more representations” (Albertoni et al., 2024). Typically represents the smallest citable unit.	Collection (if applicable)		
Collection	A repository-created grouping of datasets (may also include other research objects), generally for a discovery or administrative reason.			Datasets across a particular subject area

Table 1b: Other Key Terms Related to Granularity

Term	Description	Examples
Dimension granularity	Data being measured according to finer or coarser units of measure.	Time as measured in years, days, or seconds
Coarser	Data in relatively larger pieces.	
Component granularity	Data being subdivided or aggregated into smaller or larger components.	Single variable vs. entire data file
Custom combination	A user-created grouping of datasets or subsets thereof, across multiple datasets (e.g., multiple data sources unified by the end user).	
Dimension	The way in which scope or boundaries of an entity (e.g., dataset, collection) are described, in terms of the extent, coverage, and resolution (e.g., temporal, geographical, thematic domains).	
Entity	Level of component granularity being managed.	Collection Dataset Subset
Finer	Data in relatively smaller pieces.	
Granularity	Scale or level of detail in a set of data. Ability to represent and operate on different levels of detail in data and information.	
Instrument	Tool used to collect data.	Microscope Questionnaire Sensor
Parameter	Characteristic (e.g., value or predicate) by which a dataset can be subsetted to retrieve only the data needed.	Variables Lines of data with a certain value
Unit	The basic observable entity being measured. Entity Unit of analysis Experimental, observational, statistical, or data unit	Person

Figure 1: Component Granularity

Using the terminology defined above in Table 1b, this figure demonstrates how component granularity often involves hierarchical subsets of data that can be nested within coarser objects. Note that this generalization is not always the case, given the diversity of ways in which data can be stored (e.g., a database may be larger than a dataset). Nonetheless, this model provides a useful framework for discussing the most common hierarchical relationships.

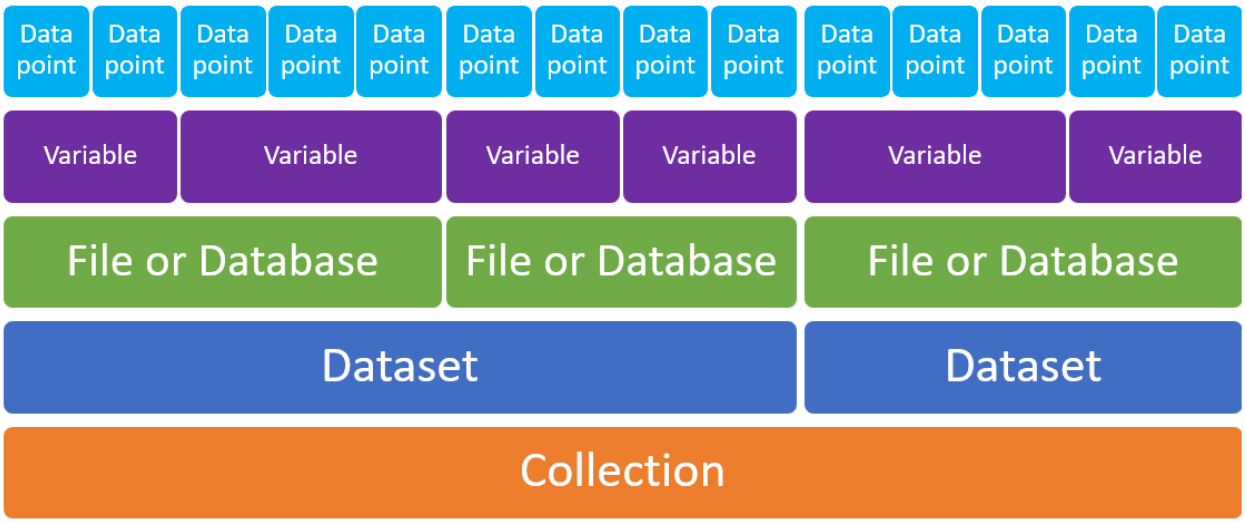
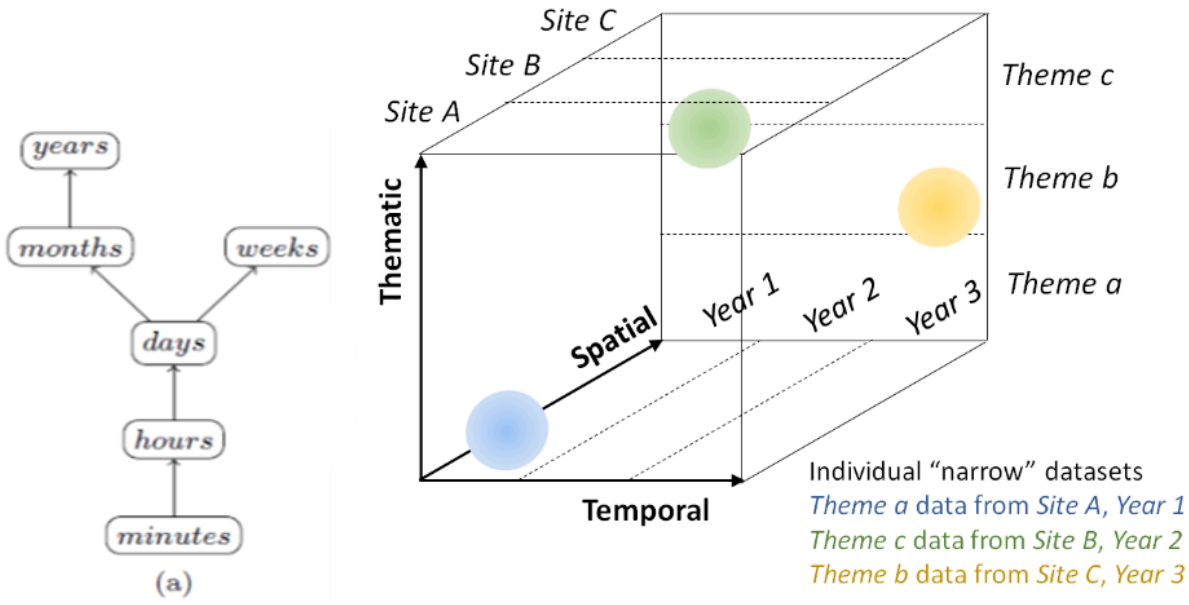


Figure 2: Dimension Granularity

This figure provides examples of ways in which granularity can be used to measure dimensions of data at a more detailed, fine-grained, and precise level. Diagram (a) on the left demonstrates how time can be measured from coarse (years) down to fine (minutes) levels of specificity, and/or parallel units of measure that do not necessarily nest into the same hierarchy (weeks not being a hierarchical subcomponent of months). Diagram (b) demonstrates scenarios in which data are processed in “chunks” according to themes (different variable sets) and sites (individual measurement stations with their own equipment), always at relevant temporal scale, before being aggregated at yearly level.

Figure 2: Dimension Granularity



In conclusion, the detailed nature of this terminology reflects the complexity of granularity issues themselves, thus reinforcing the value of this Working Group to provide guidelines to the community in this multifaceted realm, building heavily on the work of others.

Methodology

General

Building on prior RDA efforts, the WG explored key questions and collected information in order to provide guidance for data professionals on how to best support data granularity across their infrastructure and services. To produce its outputs, the main activities of the group were to:
i) produce the planned use cases (identifying the needs around this issue);
ii) conduct a survey of the community (which focused on data repositories, a key stakeholder type in the the community); and
iii) conduct an environmental scan to identify existing constraints within registries, metadata standards, and relevant existing RDA WG outputs where these data granularity concepts are applied (to ensure that the Recommendations realistically align with the ecosystem).

Three sub-groups were formed to complete each of these activities, by working independently, collaborating as a WG as a whole, and with input/feedback from the overall RDA community (such as at Plenaries). This section outlines the detailed methodology of each sub-group.

Given the integral, cross-cutting nature of granularity across a range of aspects of data services, a prime approach and value of the WG has been to leverage and build upon existing and ongoing RDA work amongst a range of groups (see Acknowledgements and Appendix 1).

Use Cases

A core of sub-group members[4] set about to create the use cases, originating from either existing documents in the field (see Appendix 1: Documents from Related RDA Groups) or members' professional experiences.[5] Given that the term "use cases" can vary in meaning, and there exist multiple ways to describe stakeholder needs regarding granularity, the group's first step was to agree on the format the use cases would take. After a review of the field the main format options considered were:

1. Narrative, open-ended descriptions of a stakeholder scenario relevant to granularity. The main benefits of this format: portrays a holistic scenario, and has the flexibility to encompass details and issues that do not fit into a consistently structured format.

2. The "user story" format: a sentence capturing a specific need of an end user, written in the following structure.
   As a [role/who], I want to [goal/what], so that [benefit to be achieved/why].
(Atlassian, 2024; Mathiak et al., 2021)

   The main benefits of this method: the ability to present user needs in a structured and discrete format; and to be able to filter use cases by certain qualities (e.g., type of role).

To evaluate these options, group members gathered a small set of user scenarios, tested writing them up in both formats, and reviewed others' tests for clarity and consistency. In order to benefit from both structure and flexibility, the group decided upon a hybrid approach: to write up use case information in a dual format, and for each scenario to include both:

1. A scenario narrative–of length between a paragraph and a page–including:
   a. An open-ended textual description of a scenario related to data granularity.
   b. The source of the use case.
   c. If applicable, an additional statement of granularity challenges highlighted by the scenario.

   and

2. One or more user story sentences (i.e., user needs) included in the scenario, added to a spreadsheet, including:
   a. A description in the format: As a [role/who], I want to [goal/what], so that [benefit to be achieved/why].
   b. Coded according to specific qualities (see Appendix 3])
   c. Mapped to its relevant narrative scenario.

[4] Listed as authors on the Supporting Output: Use Cases
[5] Given the sources, this serves as a convenience sample.

The group captured this information in a spreadsheet, available as a Supporting Output (Data Granularity Use Cases). Described in detail in Appendix 3 (Use Case Spreadsheet Documentation), the spreadsheet contains instructions for its use, a list of use cases structured according to the user story format, narrative scenario descriptions, definitions of concepts used, and instructions used by use case writers. For the most part, use case information consisted of a pairing of 1) a narrative scenario and 2) discrete user stories pulled from each narrative (e.g., generally 2-10 user stories per narrative).[6]

The use cases were written by sub-group members, utilizing both a set of standardized instructions[7] and each individual's judgment. Given the potential for differing interpretation, all entries were given a final review by an editor, and modified as needed: for clarity; expanding or de-duplicating some entries; and–to a very minor degree–coding consistency (due to the ambiguous nature of the issues, and the need to benefit from the diverse judgments of all members involved, the categories were neither prescribed nor enforced rigidly).

Survey

The primary objective of the community survey was to investigate how granularity is managed in data repositories, to explore how repositories and data services are dealing with data granularity, and what types of data they store.[8] This information can help to better describe the size and scope of repositories as well as the heterogeneous types of data granularity policies. Given this target group, the WG sought the assistance of re3data, the registry of research data repositories, leveraging their expertise and contacts (GFZ German Research Centre For Geosciences et al., 2024). The WG collaborated with re3data to design the survey, which re3data pretested. The group then modified the questions as necessary and created a final version, which re3data then distributed to the target repositories through their newsletter mailing list (see Appendix 4: Repository Survey Invitation).

The survey was comprised of three sections: demographics, entities, and policies (see Appendix 5: Repository Survey Instrument). The demographics section was designed to capture the nature of the repositories in general, including regional or disciplinary scope, the hypothesis being that clusters of repositories with similar demographics (and thus context) might have common types of behaviour regarding granularity. However, such analysis (i.e., a cross-tabulation of results by demographic variables) was not conducted due to insufficient sample size. Note that the survey also attempted to measure repository size (partly inspired by

[6] Given the flexible nature of the narratives, there necessarily is information included in the narratives not pulled out as an accompanying user story. Furthermore select cases did not lend themselves to the expected pairing (are represented by only a user story or a narrative, but not a linked pair), such as when a free-standing user story requirement was obtained from an external document without a contextual narrative.

[7] The structure and format of the use case information, and instructions for their creation, evolved over time, as the act of writing the use cases themselves led to improvements in the methods.

[8] While the original Case Statement showed an intention to survey a broader set of community stakeholders, the decision was later made to focus specifically on the data repository community.

the preliminary results of the RDA Data Repository Attributes Working Group (Witt et al., 2024), as different granularity issues can arise given the different sizes of datasets and/or level of detail (data & metadata) they contain. Yet size is a familiar concept with no standard way of operationalization. The WG arrived at a solution by documenting a general qualitative sense of repository size through three dimensions: geographic scope, thematic scope, and relative size. These dimensions were reflected in demographic questions 2.2, 2.3, 2.5, as well as by asking how many units of each entity type are in the repository (entity questions 3.4, 4.5, 5.5, 6.5 and 7.5).[9]

The entities section was designed to measure the levels of granularity the repository was using (see also Terminology), in a general exploratory manner (rather than testing a specific hypothesis related to granularity). Respondents were given the opportunity to describe up to five different types of entities (including, but not limited to, levels of granularity) and their relationships to each other. As pre-testing indicated that the term "entity" was not always immediately understood by respondents, the group added further explanatory text.

The final survey section (policy) asked about rules and decision-making practices regarding data granularity in the repository, as well as perceived challenges in granularity. Additionally, respondents were asked to provide their re3data ID, so that this information could potentially be linked to re3data metadata records at some point.

The final version of the survey was distributed through the re3data newsletter and received 47 responses (out of a universe of over 3,000 repositories). This response rate was lower than expected, attributed to a lack of awareness among potential respondents regarding the concept of granularity.

Constraints Evaluation & Environmental Scan

The purpose of this subgroup was to identify existing fundamental ways that data granularity is handled within the data infrastructure ecosystem–namely in registries, metadata standards, and relevant existing RDA WG outputs. These data granularity concepts and the WG's Recommendations must coexist and be in harmony in the settings where they are applied, to ensure that the Recommendations realistically align with the ecosystem.

The evaluation of constraints was two-fold. In one exercise, the group reviewed a set of metadata schemas and registries (see Appendix 2) in use in data infrastructure (distributed among members of the group). This list was created by team member input. The purpose of the review was to understand relevant properties of the schema (identified by group discussion), such as:

- Dimensions measured by the schema (e.g., temporal, geographical, thematic, project)

[9] Survey pretesting included questions designed to measure the size of repository by monetary value (either by the number of staff running the repository, production cost, or worth to end users). Such information was not generally or reliably known, so these questions were removed in production.

- The granularity levels and concepts (e.g., collection, dataset, subset) covered by the schema. For example, can the schema describe entities at different levels of granularity, with what supporting metadata fields, and what is their relationship to each other.

The group also reviewed a set of key related endorsed RDA WG outputs (see Appendix 1), in order to build upon, leverage, and avoid conflicts with existing RDA work. As of October 27, 2022 there were 29 endorsed outputs that were a basis for this exercise. The group evaluated RDA endorsed recommendations for any constraints, validations, or impacts related to data granularity. These results are provided below and were incorporated into the Recommendations.

The group acknowledged that, through this process, challenges in harmonization may arise. For example, whilst the group may note constraints in schemas, they may not necessarily be fixed, as improvements may be requested to enhance these schemas via their various governance structures as an outcome of the recommendations. It may be possible that Recommendations, in order to be the most robust, may conflict with existing RDA Outputs; in that event, a rationale will be provided.

Findings

Use Cases

The documented use cases (Supporting Output: Data Granularity Use Cases) provide a rich, qualitative set of data to inform this report's recommendations and readers' understanding of data granularity in their contexts. The use cases were coded to indicate the relevance according to key characteristics: user roles, levels of granularity, and categories (type of use) (see Appendix 3 and Use Cases output for full documentation of categories). An aggregate analysis of the use cases yielded descriptive statistics among these characteristics, including frequency and relationships amongst different characteristics (cross-tabulations). These results provide insights and an indication of significant granularity issues to be addressed.

Frequency statistics

A total of 64 use cases were recorded (from narratives yielding multiple use cases each) with the following key characteristics:

- Role
    - 28 data consumer
    - 22 repository
    - 10 data producer
    - 2 data manager/steward/curator
    - 1 funder
    - 1 research administrator

- Level of granularity
    - 19 custom subset
    - 18 dataset
    - 8 variable
    - 6 most/all [of these levels involved]
    - 4 file
    - 4 collection (of datasets)
    - 4 custom combination (among different data sources)
    - 1 instrument (of data collection)

- Category/Type
    - 19 findability/discovery
    - 16 curation/deposit/preservation
    - 12 identification/citation/metrics
    - 9 access
    - 4 analysis
    - 3 other (2 relevance assessment, 1 public communications)
    - 1 interoperability

Of the 64 use cases collected, nearly half came from the perspective of a data consumer, another third as a repository, approximately one seventh as a data producer, and a handful from the other perspectives. One can identify major themes and categorize key types of use cases for these roles. However, it is important to note that–due to the project design–the findings are neither exhaustive nor necessarily representative (e.g., the limited number of perspectives from other role types, for example, may indicate a bias in the method of convenience sampling).

Various levels of granularity were described in the use cases, with the most frequently referenced by far being custom subset and dataset, whereas instrument ranks the lowest. This may imply that when it comes to data reuse (from data consumer perspective) or data curation (from repository perspective), subsets of data often are involved. The few relationships to custom combination could possibly indicate that fewer data granularity issues arising in that process; or this may be due to sampling, as one would expect data granularity issues to potentially arise very frequently when integrating data from different sources (given challenges in merging data at comparable levels of granularity, which may even more complex than the custom subset process (i.e., extracting slices of data from a whole)).

All of the categories (stages of the data lifecycle) came up among the data granularity issues, with the most frequently identified being: findability/discovery, curation/deposit/preservation, and identification/citation/metrics. This reflects the stages at which granularity issues most frequently arise within this sample. Cross-tabulations were performed on the data to observe potential trends among subsets, demonstrating relationships between ‘Type of roles’ and either ‘Category/Type of use cases’ or ‘Level of granularity.’

Table 2. Type of use case (Category) broken down by role

Type of use case	**Data consumer**	**Repository**	**Data producer**	**Data manager/ steward/ curator**	**Funder**	**Research administrator**	**Total**
Findability/ discovery	15	4	0	0	0	0	19
Access	6	3	0	0	0	0	9
Interoperability	0	0	1	0	0	0	1
Analysis	4	0	0	0	0	0	4
Identification/ citation/metrics	1	6	3	0	1	1	12
Curation/ deposit/ preservation	0	8	6	2	0	0	16
Other	2	1	0	0	0	0	3
Total	**28**	**22**	**10**	**2**	**1**	**1**	**64**

Table 3. Level of granularity broken down by role

Level of granularity	**Data consumer**	**Repository**	**Data producer**	**Data manager/ steward/ curator**	**Funder**	**Research administrator**	**Total**
Most/all	0	2	2	2	0	0	6
Collection	4	0	0	0	0	0	4
Custom combination	1	0	3	0	0	0	4
Custom subset	5	13	1	0	0	0	19
Dataset	8	5	3	0	1	1	18
File	3	0	1	0	0	0	4
Variable	6	2	0	0	0	0	8
Instrument	1	0	0	0	0	0	1
Total	**28**	**22**	**10**	**2**	**1**	**1**	**64**

From the perspective of data consumers, there are a range of tasks where data granularity impacts the ability of the user to succeed. By far, most commonly these scenarios relate to finding relevant data, with granularity arising as a factor in challenges such as: understanding granularity in the description of the data, identifying what types of data are available, using criteria to filter a search, or finding different/specific versions of a dataset. Within the use cases,

the principal issues of granularity for those seeking and reusing data relate to the ability to find the right dataset and for it to be effectively analyzed. In addition to discovery, user tasks include downloading the specific data required (rather than a wider dataset), working with data across multiple datasets based on a common variable, and interrogating data within the repository (e.g., preview or visualization before downloading). Other challenges include citation of reused datasets and locating relevant code.

While the data consumer tasks are more often one-off or time-limited activities, the granularity use cases identified for repositories tend to be ongoing management activities. These include: providing robust SEO (search engine optimization), reporting on data usage across different levels of granularity, and reporting on the size of the repository. Many repository tasks are the functional equivalent of achieving data consumers' goals.[10] This speaks to a need to embed granularity at a systemic level, so that the platform provider can achieve the use cases of the data consumer.

For data producers, the most common category of use cases regarded curation/deposit/preservation. Examining data producers' primary goals and motivations, issues of data granularity occur in the context of practical requirements for research (e.g., depositing data to comply with a journal data policy). Regardless of their motivations for sharing data (whether a requirement or altruism), data producers must manage versions of their data in order to provide an accurate and appropriately described set of data. Given this analysis, when considering the needs of data producers, data repositories should develop features to support the management of versions of data in a way that addresses data granularity issues. Table 4 shows relationships between 'Level of granularity' and 'Category/Type' of use case.

Table 4. Level of granularity broken down by Type of use case (Category)

Level of Granularity	**Findability/ discovery**	**Access**	**Interop- erability**	**Analysis**	**Identifi- cation/ citation/ metrics**	**Curation/ deposit/ preservation**	**Other**	**Total**
Most/all	1	0	0	0	0	4	1	**6**
Collection	3	0	0	1	0	0	0	**4**
Custom combination	0	0	1	1	1	1	0	**4**
Custom subset	1	5	0	0	5	8	0	**19**
Dataset	9	1	0	0	6	2	0	**18**
File	1	0	0	0	0	1	2	**4**

[10] E.g., "Provide a range of query interfaces to accommodate various data search behaviors." "Provide multiple access points to find data (e.g. search, subject browse, faceted browse/filtering)." "Provide an API / machine actionable landing page to access metadata and data via query re-execution."

Level of Granularity	Findability/ discovery	Access	Interop-erability	Analysis	Identifi-cation/ citation/ metrics	Curation/ deposit/ preservation	Other	Total
Variable	4	2	0	2	0	0	0	**8**
Instrument	0	1	0	0	0	0	0	**1**
Total	**19**	**9**	**1**	**4**	**12**	**16**	**3**	**64**

Based on the data granularity use cases compiled, the custom subset of data represents the most frequently identified level of granularity, with issues across three types of data granularity scenarios: access, identification/citation/metrics, and curation/deposit/preservation. Thus the need to create, manage, access, and cite custom subsets poses important data granularity issues for all stakeholders involved, prominently data consumers, data producers, and repositories. Another finding worth highlighting is that the need for findability/discovery arises frequently at the dataset level of granularity, and also for collections and variables.

Challenges arising

The use cases document a number of challenges regarding data granularity (some highlighted in a dedicated column “User Problems Identified”). From the perspective of data discovery, it is uncertain how often metadata fields account for (and thus are searchable by) levels of data granularity, and how well repository search interfaces are designed to accommodate data granularity.

Important is the challenge of uniquely identifying data at a particular granularity, such as the common task to “uniquely identify custom subsets generated by queries.” While some repositories do provide PIDs for a result based on a search query,[11] more commonly repositories apply a PID only at the level of dataset granularity. The challenge of identifying custom subsets (a demonstrated use case) therefore is also tied to a challenge in PID granularity, relevant to both data producers, infrastructure providers and data consumers. Several groups in the field have been working to better understand and address this issue, such as the RDA Working Group on Dynamic Data Citation (Rauber et al., 2020).

Datasets are not always adequately linked to other research outputs or related metadata; from a granularity perspective a dataset could underpin several research papers, or several datasets could underpin a single research paper, hindering the potential for comparative analysis.

As access to research data plays a key role in reproducibility, so the granularity of a dataset must be comprehensible in reuse in the context of other outputs like the research publication, code and protocol. In particular where studies draw on parts of existing datasets, the ability to

[11] Such as the Global Biodiversity Information Facility (Global Biodiversity Information Facility, 2024).

reproduce the data depends on consistent granularity. If a secondary analysis simply cites existing datasets as a whole for example, rather than the parts reused, this may be insufficient to enable reproducibility and thus confirm the integrity of the results and analysis.

A similar problem applies to using multiple datasets for comparative analysis. To answer many research questions, researchers often must combine multiple datasets, according to common variables. This relies on the datasets being discoverable and available with common variables, with the ability to combine and harmonize data at comparable levels of granularity. This scenario also raises challenges in how best to cite a custom combination dataset (with data from across multiple sources).

In summary, the use cases demonstrate important themes to be addressed, among the range of needs for data granularity in research.

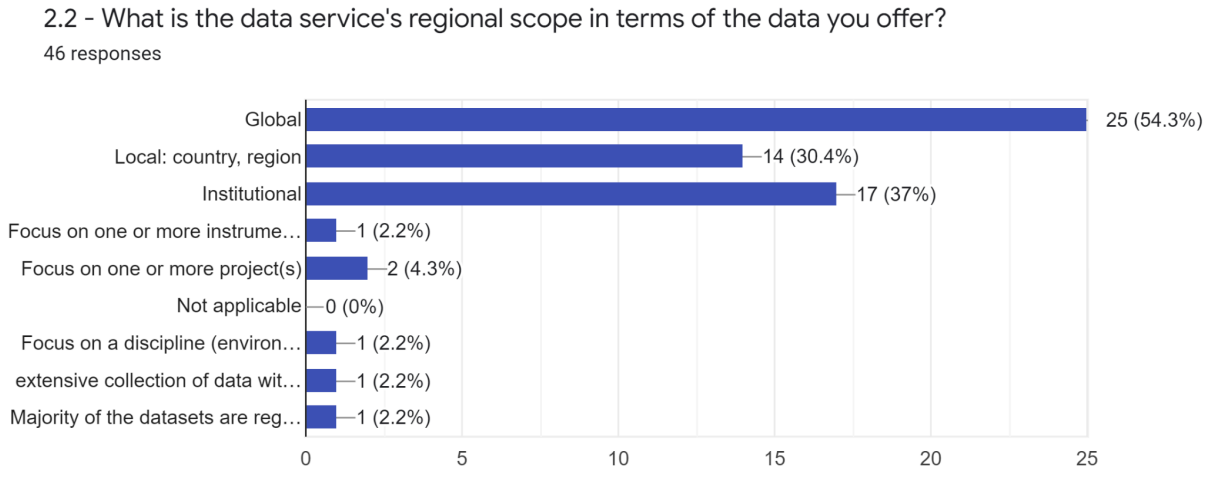
Survey

The survey of data repositories yielded a number of interesting findings regarding the general characteristics of responding repositories and how they view issues of data granularity.[12]

Demographics

The survey shows the breadth among responding repositories, regarding both geographic (Figure 1) and disciplinary (Figure 2) scope.

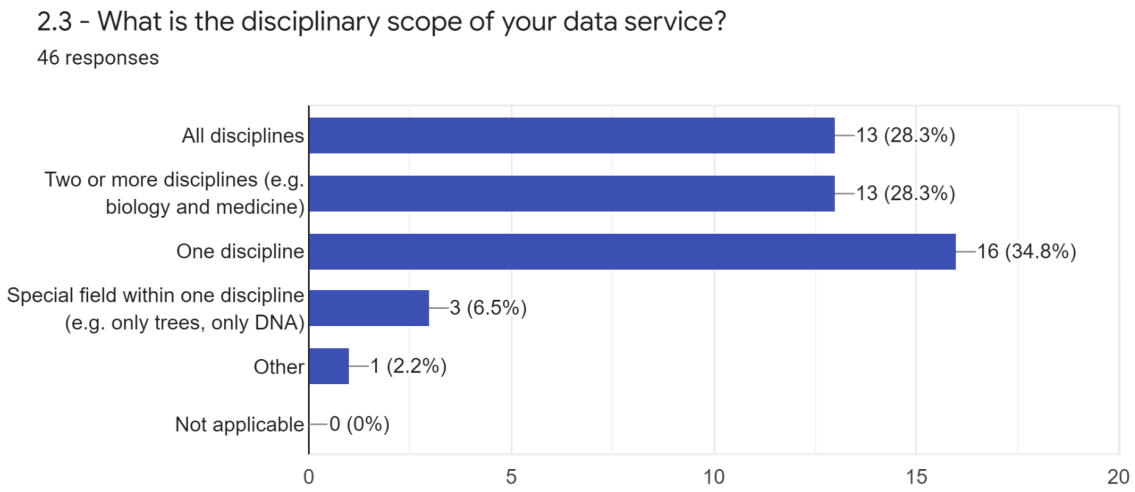
Figure 1: Geographic Scope



Most repositories indicated a global collections scope, with significant portions having data covering a specific country, region, or institution.

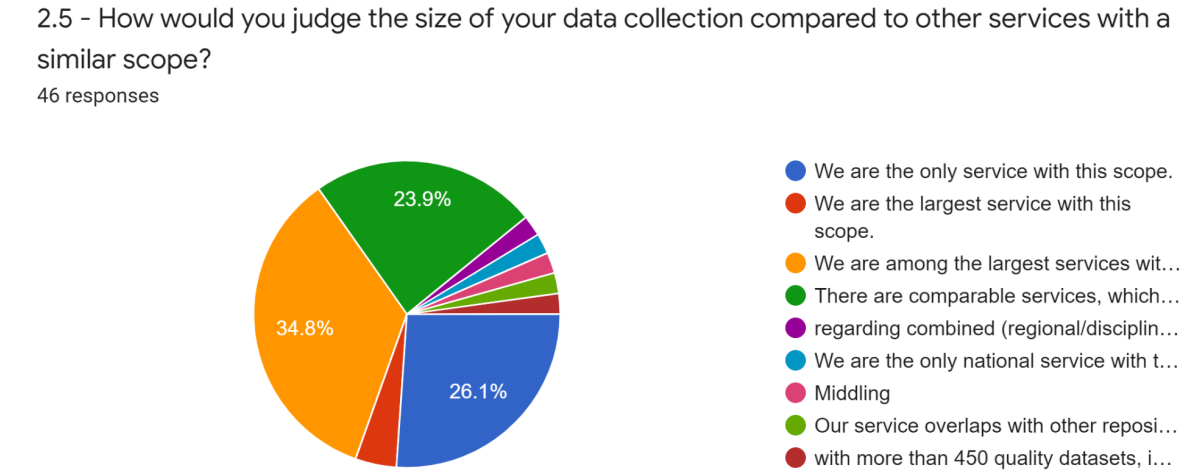
[12] For the full text of the questions and response categories, see Appendix 5 (text is truncated in figures below).

Figure 2: Disciplinary Scope



Most of the services in our sample have a large scope, with multiple data providers. Figure 3 demonstrates that while it is most common for the repositories to cover one specific discipline, many cover multiple fields. Repositories covering specific sub-fields are relatively rare.

Figure 3: Relative Size Estimates



When comparing the size of one's repository to others' with similar scope, approximately one third judge themselves to be among the largest services with this scope, one quarter are the only service with this scope, and one quarter judge themselves to be smaller than average. Repositories indicating that they are "the largest service" were very rare.

The survey also attempted to collect data about size in absolute terms. For each type of entity described (see section below), the respondent was asked the open-ended question: “How many units of this entity type are in your data service?” However, the content and format of responses varied widely (e.g., ”hundreds”, “~115 TB”, “1000+”, “182”, “80?”, “5 billion+ individual data points (sequence annotations)”, “170798”, etc.). Due to the diversity of how repositories measure size, and uneven terminology in this field (i.e., not all respondents may have had the same understanding of “datasets"), the Working Group realized it was not feasible to utilize these open-ended responses to develop an absolute quantitative measure. Furthermore, it is uncertain the degree to which the sample was representative (e.g., Were small and specialized repositories sufficiently represented? Were interdisciplinary repositories overrepresented?).

The four dimensions of size (geographic scope, thematic scope, relative size, plus the estimated numbers of entities) seem to be only weakly correlated, if at all. The fact that one quarter of the repositories consider themselves to be the only service with a given scope is maybe not as surprising as it appears to be. Some have unique mission statements (e.g., “vascular plants in Bavaria”), some are unique through a combination of regional and disciplinary scope, while others are the dedicated data repository for their institution. While the comparison dimension seems like the best indicator of them all, the authors deem it to be unreliable, due to the vagary of what scope means in this particular context. Nonetheless, these dimensions[13] give a general sense of the magnitude of repositories and their own sense of their size, which provides a context for how these repositories manage granularity.

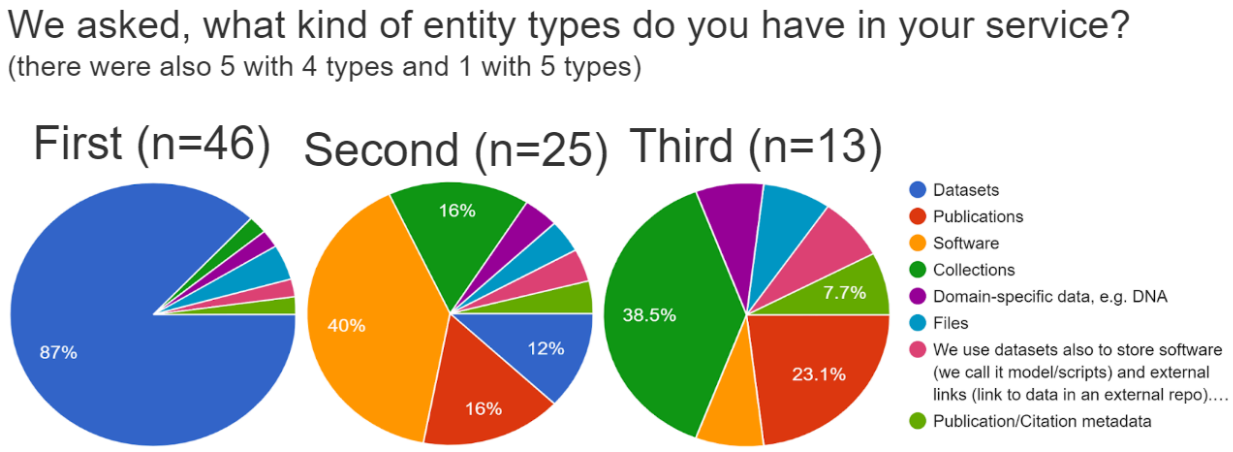
Entities

The responses to these questions show the types of entities repositories possess, and how they think of them. For each entity described, the respondent was first asked to indicate entity type, meaning the types[14] of information stored (levels of granularity were asked in a subsequent question). While respondents were able to describe up to five types of entities, most described only one or two.

[13] In hindsight, the survey could have included yet another dimension of size–personnel (i.e., by asking about the number of staff involved in running the data repository). While also influenced by the nature of the data and the types of curation performed, such responses might have given an idea of the involvement of staff in curating this data, a foundation for understanding how much granularity issues affect them directly.

[14] It is important to note that these type categories are not mutually exclusive; for example, one could store “domain-specific data” in the form of “files” and also as part of “datasets” within collections.

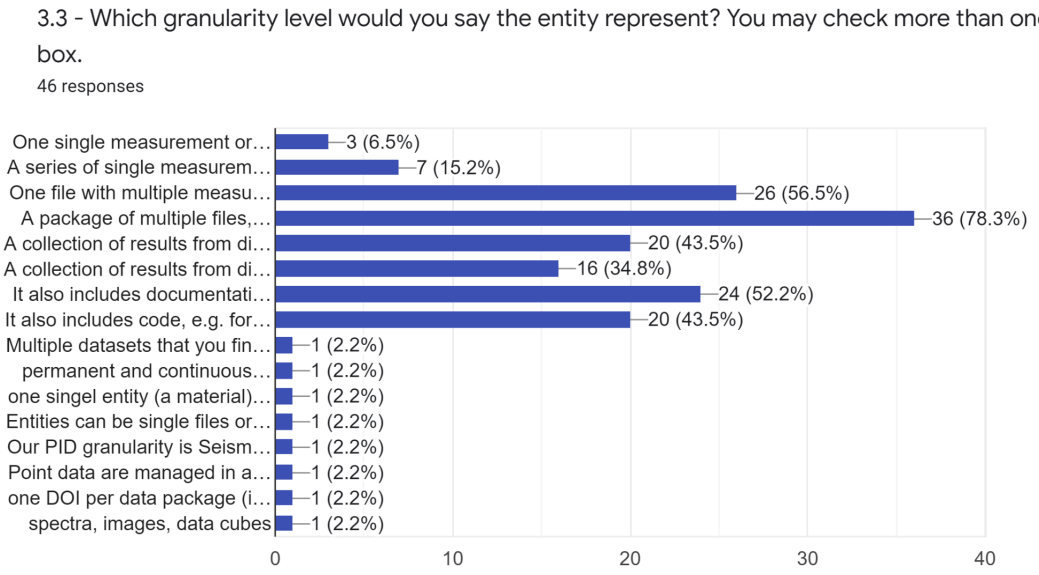
Figure 4: Entity types



As shown in Figure 4, the vast majority of respondents first reported storing datasets; many went on to share information on other information types like software and publications, as well as storage mechanisms like files and collections.

When asked about the levels[15] of granularity represented, respondents were able to pick more than one among a list of several.

Figure 5: Entity granularities

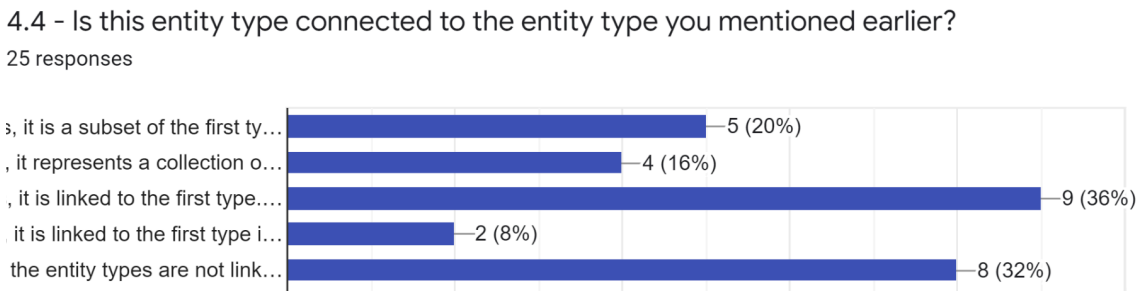


As seen above, various groupings of data (e.g., a set of data points, one or more files or collections) were the most common, but the long tail of different granularity types is notable.

[15] Note that this list (an experimental set for the survey) is not the same as that listed earlier under Terminology (a focused set refined over the course of the project).

As seen in Figure 6 below, many–albeit not all–entities are related to each other, whether it be hierarchically (first two lines) or through some other relationship (lines 3 and 4, as one might find, for example, between data and a related publication).

Figure 6: Entity relationships



As discussed earlier, respondents may not have had the same terminology meanings in their mind when answering these questions (e.g., the same understanding of what is a dataset, subset, etc.). The results showed an interesting variety, and the act of having to answer these questions may have triggered respondents to think more deeply about the different levels of granularity in their repositories.

Policies

Collection policies with regard to granularity provide another window on how repositories work and plan for granularity.

Figure 7: Policy guidelines

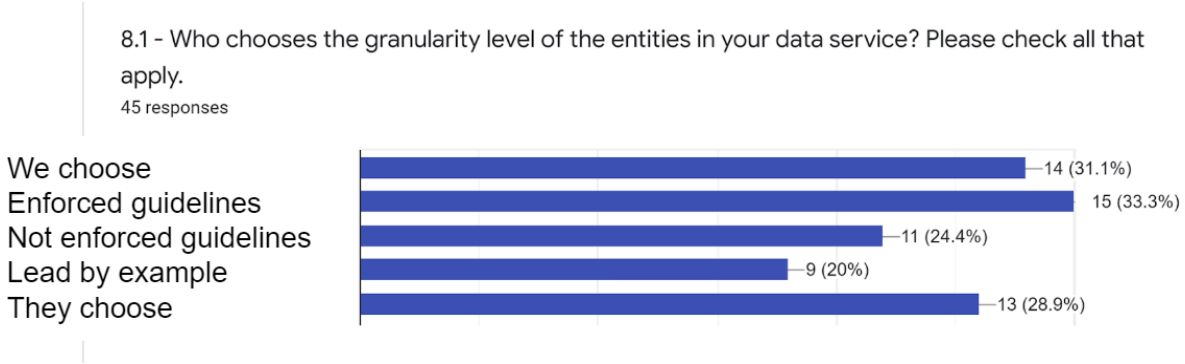
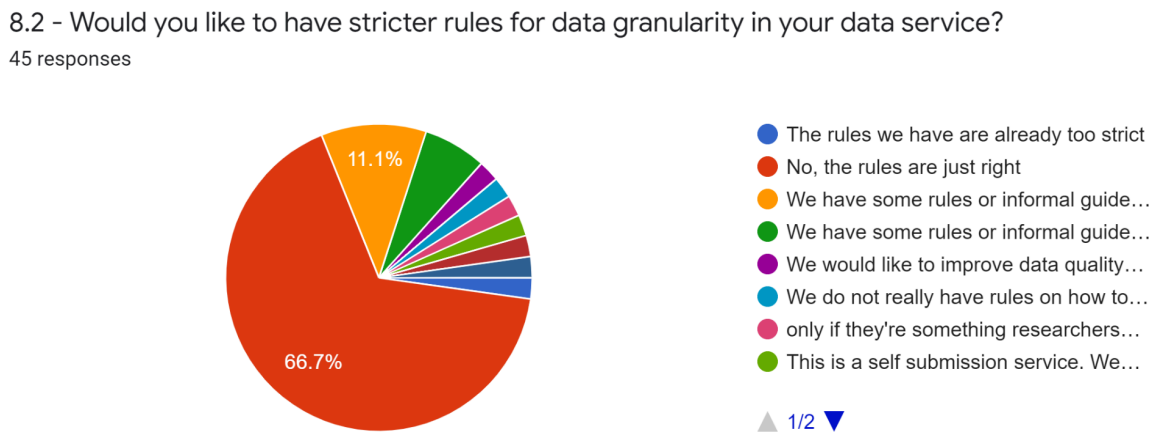


Figure 7 demonstrates that there is no leading approach for the person/body responsible for choosing/authorizing the granularity of data that may be collected. Often repositories will achieve a desired level of granularity through guidelines or repackaging the data, but a laissez faire approach also is common.

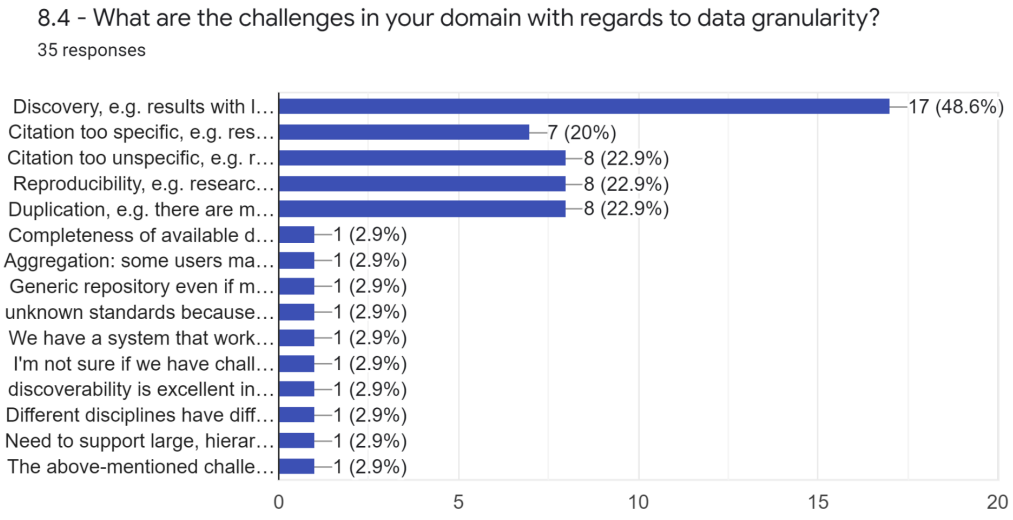
Figure 8: Policy satisfaction



When asked, two thirds of respondents reported being satisfied with their existing granularity rules, yet the other third saw the need for improvement (with 11% desiring stricter rules).

Finally, the survey asked about the leading challenges in data granularity.

Figure 9: Granularity Challenges



All challenges listed were cited in some context, with many additional “other” challenges (those above with a single response). Different types of citation challenges were widespread, but leading was the challenge that granularity can pose for users being able to discover the type of data that they need. In general, while the survey showed a range of approaches to granularity, it identified common situations and challenges which informed this report and its recommendations.

Constraints Evaluation & Environmental Scan

This review of metadata schemas and endorsed RDA outputs demonstrate a need for more comprehensive, consistent and structured metadata about granularity to serve community recommendations and use cases. Furthermore, they present a key set of constraints within which stakeholders need to operate when implementing their data granularity approaches.

Metadata schemas

Metadata schemas offer widely varying possibilities and generally inconsistent metadata for expressing granularity information, in terms of both component and dimension granularity and the relationships between granularity levels. Table 5 summarizes observations for these metadata with some representative examples.

Table 5: Observations on granularity concepts found in widely used metadata schemas.

Metadata facet	**Observations and examples**
Temporal dimension	Temporal extent metadata are typically supported, although the format to represent date and time is not always specified by the schema. Temporal resolution or regularity of data is rarely supported in any systematic way. Examples: ● DataCite: Date element with dateType Collected option for the time range for which the data were collected. DateInformation can give more free text details. ● Dublin Core: coverage has temporal subproperty to indicate named period, date or date range. ● ISO 19115: EX_Extent class contains TemporalExtent metadata that can express time range or instant.
Geographical dimensions	Points, bounding boxes and polygons usually have explicit guidance for representing the geographic extent, but the resolution is typically not represented. Datum information and altitude or depth do not always have specific metadata fields allocated. Placename and feature type information are also often possible, but not well-standardized. Examples: ● DataCite: Geolocation includes latitude/longitude defined by point, box, polygon and/or description. There is no altitude/depth option. ● Dublin Core: coverage has spatial sub-property for geospatial information like location, jurisdiction or geographic coordinates (point or box). ● ISO 19115: EX_Extent class that contains metadata for GeographicExtent (latitude/longitude) via points or polygons, and VerticalExtent (altitude/depth) range.
Thematic	Often schemas support thematic areas with keyword selection, but

Metadata facet	Observations and examples
dimensions	ontologies may vary or not even be used. In some cases, thematic information is further outlined with metadata for variables. Examples: ● DataCite: subject field with classificationCode option ● DublinCore: subject with recommended vocabularies ● EML: keywords that can be associated to ontologies
Project dimensions (e.g., project, expedition, experiment)	While many repositories and researchers organize data deposits in alignment with a research project or data collection event, this is typically only described within title or abstract metadata. Few schemas have fields that explicitly capture this information. Examples: ● ISO 19115-2 provides acquisition details like the operations, platforms and instruments that were used to acquire the data.
Relationship frameworks	The ability to represent relationships between levels of granularity varies greatly, with some only allowing hierarchical relationships. Recursiveness or nesting is not always supported. Another common limitation is that only the repository can curate the relationship metadata, rather than providing end-users tools to define collections or subsets themselves. Examples: ● DataCite: ResourceType options include dataset and collection. Entities can have relatedIdentifiers with relationTypes IsPartOf and HasPart ● ISO 19115 MD_Metadata class includes a parentMetadata field which enables an association to a higher level collection. Each record also indicates the 'resourceScope' which includes dataset , collection, series, aggregate and more. ● EML has a hierarchical structure which handles many relationships; one can use many data table sections within a single file that describes a single project, and each data table section may have many variables described.
Data size metadata	The ability to express data size is inconsistent, being expressed in terms of data volume, number of data points, number of files, etc. Information about how size scales for subsets or collections is typically not indicated. Examples: ● DataCite: Size field is free text, yielding highly varying information. ● Dublin Core: extent field for size and duration, recommending file size in megabytes ● ISO 19115: MD_Distribution class has a transferSize field expressed in megabytes. ● EML: Tags support size/volume reporting.

Metadata facet	Observations and examples
Format metadata	Most schemas allow for representation of the types of data formats included, although consistency as to how it is described is typically not enforced. Examples: ● DataCite: A format field exists, but is free text so this can be highly varying information. ● ISO 19115: MD_Distribution class has a distributionFormat field. ● EML has a physical module to describe file attributes including format.

Endorsed RDA Outputs

Numerous outputs recognized the need for better consistency and clarity of data granularity concepts and applications. The following table summarizes key findings extracted from the review of RDA endorsed outputs (see Appendix 1 for detailed list with full citations). Only those outputs which have notable influences are presented here.

Table 6: Core findings that express data granularity concepts or dependencies from key RDA endorsed inputs

Group	Key findings
Data Citation WG	These recommendations provide a scalable mechanism that allows the precise, machine-actionable identification of sub selections of data, with time-stamped queries preserved in and retrievable from a query store with query PIDs. Data versioning should be supported with new persistent identifiers assigned.
Data Discovery Paradigms IG	Granularity was discussed as an important aspect of required metadata. Recommendation 3 indicates that granularity influences users' judgment of data fitness for purpose.
Data Foundation and Terminology WG	While this output confirms that there is a need for greater clarity for collections, it outlines a number of characteristics and conclusions of relevance: ● “A digital collection is an aggregation which contains digital objects and DEs [data elements]. The collection is identified by a PID and described by metadata.” ● Recursive collection (e.g., collections with collections) examples are provided. ● “Important in all solutions is that schema, element semantics and relational semantics are made explicit to enable machine based interpretation.”
Data Type Registries WG	This output includes the notion that the data type (where a data type is a characterization of the data structure, context, and assumptions) is

Group	Key findings
	applicable at multiple levels of granularity, from individual data points up to large data sets. Thus, the metadata to describe the data type should be applied to all levels of granularity. The group recommended having a data type registry with PIDs to include this information.
Data Usage Metrics WG	The issue of data granularity is embedded but not fully addressed in this guidance. For instance, the report states: “Download volume (i.e., file size) can be reported. There are widely varying practices in the research data community regarding the granularity and structure of datasets, components, and collections. Reporting download volume makes it easier to compare usage for research data packaged into datasets with different granularity.” This approach may have some validity within a repository or discipline with the same type of data, but is less suitable when comparing across data types. Trends may be more useful than the absolute volume. Of note, the Group also suggests having the granularity options for the reporting time period, meaning that access or usage tracking would need temporal information.
Data Versioning WG/IG	In this output, ● Issue 6 relates to granularity citability: “What level of granularity is appropriate for Persistent Identifiers (PIDs)? The granularity of PID: Should every revision receive a PID or each release/certain level of revision receive a PID?” ● Principle 3 focuses on granularity: “Identification of Data Collections (Granularity): A collection of data may be the result of successively generated datasets. The full set of aggregated data (data collection) can be seen as ‘works of works’, and may be organized in a number of sub-collections to be served by a data repository or archive. The collection of works must be identified and versioned, and so shall be its constituent datasets or individual works. This practice of identifying elements of a collection, and identifying the collection as a whole, is similar to the established bibliographic practice of identifying individual articles in a journal and identifying the journal series as a whole. The granularity is to be determined by the use case to provide a way (or ways) of identifying parts and versions whenever the practical need arises. Entire time series should be identified as collections, as should be time-stamped revisions, if the series is updated frequently.”
Metadata Standards for Attribution of Physical and Digital Collections Stewardship	This output contains a list of systems for tracking and aggregating metrics for diverse types of research outputs, including data. It also advocates for infrastructure to link specific digital objects with curatorial agents and activities.

Group	Key findings
Publishing Data Services WG	This work relates to the notions of digital object relationships, pointing out that these relationships are added over time. Thus, collections may be introduced as new research objects over time as new related outputs are produced, or given persistent identifiers to facilitate the relationships. The Group also indicates the notion of an assertion date at which the relationship is introduced. The Assertion Date, Assertion Source and Research Object PID are what are deemed essential for the relationship metadata.

See Appendix 1 for detailed list with full citations

Overall Trends and Observations

Examining these findings together illustrates several key overarching trends and observations related to data granularity.

- There remains a need to embed data granularity at a systemic level within data infrastructure, in the way that supports the (often aligned) needs of the stakeholder communities.
- The dataset still often operates as a primary level of data granularity.
- Custom subsets are key to many use cases, but are a level of granularity still requiring additional support in the field.
- Granularity in data discovery remains a prominent challenge.
- Levels of granularity within metadata are valuable, but not always sufficiently present.
- Levels of granularity are interrelated (often–but not always–hierarchically, with potential to change over time), yet methods to represent those relationships may be insufficient.
- In general, there exists the need for better consistency and clarity of data granularity concepts and applications.
- Lastly, tangential to granularity levels per se, the ability to express data size is inconsistent.

Recommendations

Given these results, the Working Group provides the following set of recommendations for stakeholders to optimize support of data granularity. Whilst these recommendations will not be universally feasible, and should be adapted as appropriate, they nonetheless outline a common set of ambitions for the community to work toward, both individually and collectively.

Recommendation 1, Entity Characteristics: Entities should possess a key set of characteristics and relationship properties.
Affected stakeholders: repositories, publishers, data producers
Detailed description:
Repositories should manage entities (e.g., collections, datasets, subsets) to have the following characteristics and relationship properties:

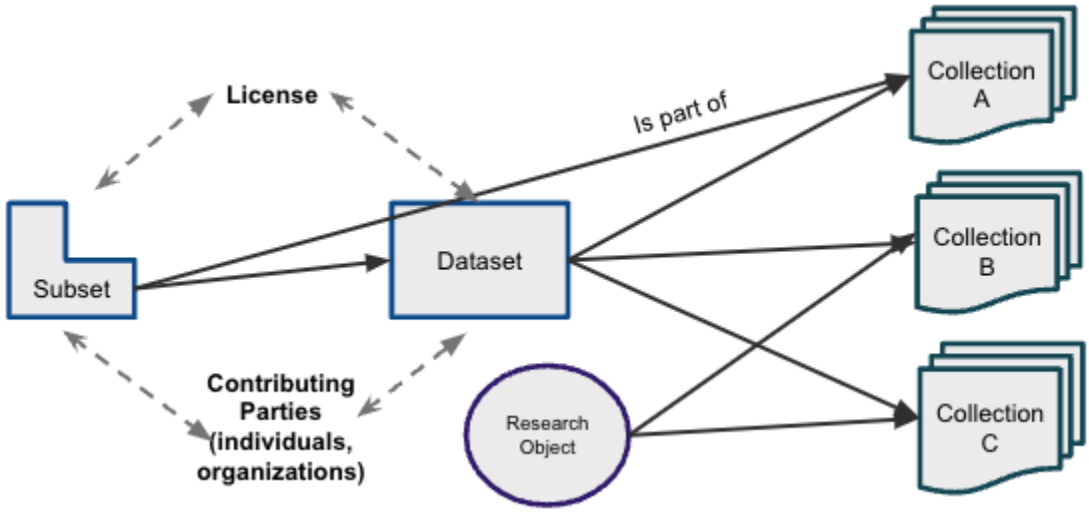
a. Dimension metadata should be structurally embedded in the dataset as a means by which the dataset could be sliced and diced into subsets.
b. Datasets should give credit to producers (aka authors, contributing individuals, and organizations), both in the metadata and under a license.
c. Collections may include collections, datasets, subsets, or other research objects.
d. Collection licenses should be optional, and should not override any of the licenses of parts of the collection.
e. Research objects of a collection may change over time, so long as there is a record of when the relationship was introduced.
f. A dataset or subset can belong to any number (0 to N) of collections.
g. Subsets inherit the license and credited parties of the datasets.
h. For infrastructure employing DOIs, employ DataCite Event Data (DataCite, 2024) to express relationships among entities (e.g., collections, datasets, subsets).

Limitations:

- Many schemas only support hierarchical collections, preventing implementation of 1f. In such cases, it may be possible to indicate that an object is related to another collection.
- The nature of some data formats may prevent embedding of dimension metadata.
- Some datasets may have multiple licenses, complicating the ability to have a clear derived license for a subset.

Figure 10 provides a visual example of how these entity properties are interrelated.

Figure 10: Entity Relationships



Datasets have defined standardized dimensions (e.g., temporal, geospatial, thematic) and boundaries within those dimensions defined with corresponding resolutions. Subsets may occur based on or within those dimensional boundaries. Collections may also describe dimensions, but perhaps may be defined more loosely.

Recommendation 2, Metadata: Infrastructure should make use of metadata to support and enable documentation of (and subsequent services based on) data granularity.
Affected stakeholders: repositories, publishers
Detailed description: Infrastructure should make use of robust and interoperable metadata to support and inform data consumers regarding data granularity, on the levels of, including:

a. Existing metadata schemas should include sufficient (mandatory or strongly recommended) fields for data granularity, involving clear, consistent and interoperable ways to populate these fields. The semantics/vocabularies may differ based on discipline at a more precise level but general dimensions should be defined at an interdisciplinary level.
b. Stakeholders should document granularity type and relationships, in particular hierarchical relationships, as part of the metadata scheme.
c. Dimension or component metadata fields should be clear, consistent, comprehensive and interoperable (e.g., appropriate fields with controlled vocabularies and units of measure to the greatest extent).

Limitations: Given that data producers may use varying granularity terminology, it may be difficult to readily produce consistent granularity metadata across datasets.

Recommendation 3, Data Policy: Policies should be put in place to support data granularity.
Affected stakeholders: repositories, publishers, funders, research institutions
Detailed description: Stakeholder organizations should put in place policies to support granularity, including:

a. Repositories, institutions, funders, and/or publishers should specify granularity

requirements in data policies, data guidance, and management documents (e.g., data management plans) whenever possible.

b. Repositories and/or disciplines should define consistent notions of dimensions and segmentation for datasets to better facilitate workflows and comparable metrics.

Limitations: Policies will be less impactful if not accompanied by appropriate guidance, resources for compliance, and a common understanding regarding granularity terminology across stakeholder types.

Recommendation 4, Discovery & Access: Infrastructure providers should enact ways to enhance discovery and access at multiple levels of granularity.
Affected stakeholders: repositories, publishers, data consumers
Detailed description: Infrastructure providers should enact ways to enhance discovery and access at multiple levels of granularity, including:

a. Leverage the dimension metadata for discovery via filtering and sorting features.
b. Discovery systems should leverage (i.e., index) metadata at multiple levels of granularity (especially those containing detailed information), tailored to the needs and practices of their user community.
c. Custom subsets should be able to be consistently identified and represented in a query store that records dimensions that characterize the subset.

Limitations: Adaptations may need to be made to ensure that more granular discovery mechanisms align with the search behavior of the user communities.

Recommendation 5, Identification & Citation: Infrastructure should put into place mechanisms that support the identification, citation, and tracking of entities at multiple levels of granularity.[16]
Affected stakeholders: repositories, publishers, data producers
Detailed description: Infrastructure should put into place mechanisms that support the citation and tracking of entities at multiple levels of granularity, including:

a. Multiple levels of granularity (i.e., collection, dataset, custom subset), all but the finest, should be independently identifiable and citable with a persistent identifier (PIDs) and resolvable landing page. At a conceptual level, the *DCC How-to Guide* 'How to Cite Datasets and Link to Publications' recommends "that repositories assign identifiers at the finest level of granularity at which the data can be said to form an intellectual whole." (Ball & Duke, 2015) Note that specific guidelines and RDA outputs exist for identifying dynamic subsets.(Rauber et al., 2020, 2021)
b. Citation and usage tracking systems should enable:
    i. Passing on or inheriting information to finer or coarser levels of granularity as appropriate. For example citing–or tracking usage of–a dataset should also count as a citation on the collection itself, and vice versa; and

[16] Recommendation 5 draws heavily on a couple of key publications in the field regarding data citation (Ball & Duke, 2015; Rauber et al., 2020, 2021).

ii. Aggregation and/or hiding of finer-grained citations where presenting a single, coarser citation would be most practical (e.g., populate a creator's RIS or ORCID record solely with the citation for the dataset, rather than one for each file) (Dataverse Users Community, 2022).

c. When citing data, data consumers should cite "the finest-grained level [available from the repository] that meets the need of the citation" (Ball & Duke, 2015).

Limitations:

- Adaptations may need to be made to ensure that more granular citation mechanisms align with the behavior of the user communities.
- Given the financial cost of creating PIDs, organizations may need to balance available resources with their desire to add PIDs at multiple levels of granularity.
- Issues of scale may make the persistent identification of finer-grained objects unfeasible in certain contexts (e.g., if a dataset has hundreds of files or millions of subsets).

Recommendation 6, Examination: Stakeholders should proactively examine and assess how best to support data granularity in their context on a periodic basis.
Affected stakeholders: all
Priority Use Cases supported: all
Detailed description: Optimal granularity approaches depend upon various contexts and evolving factors. Given that, stakeholders should proactively and continually examine and assess how to best support data granularity in their context. They may do so by asking themselves key questions around granularity, such as:

a. What are the levels of granularity used by our community?
b. In what ways can the data be subdivided?
c. What are the dimensions that define the boundaries of the dataset?
d. Do the metadata used adequately describe the dimensions, subset options, and other aspects of granularity? To what extent are these metadata interoperable with other systems?

Limitations: Whilst asking such questions will be valuable, it yields ambiguous results. If the optimal granularity strategy is unclear, stakeholders may need to make trade-offs and best guesses in order to make decisions and move forward.

Recommendation 7, Terminology: Stakeholders should adopt the terminology recommended in this document, in order to facilitate common understanding and interoperability of granularity.
Affected stakeholders: all
Priority Use Cases supported: all
Detailed description: Stakeholders should adopt the terminology recommended in this document, in order to facilitate common understanding and interoperability of granularity. In cases where there is a specific reason to use alternate language, documentation should include a mapping to standard terminology.
Limitations: Terminology standardization may need to be balanced with the practices of the specific user community.

Conclusion

As a result of its work (e.g., RDA plenary discussions, working group meetings, sub-group projects), the RDA Data Granularity WG landed at a set of high-level recommendations that recognizes the diversity of research data and disciplinary norms. While members initially anticipated recommending a more prescriptive approach to data granularity, the group ultimately realized that was not possible or even necessary due to dataset diversity and shifting expectations for research data. Furthermore, given evolving norms for research assessment, the field is seeing less emphasis on data usage metrics based on citations, and more emphasis on data usage narratives and relationships, partly due to the challenge of interpreting citation metrics when the underlying datasets have inconsistent granularity. Moreover, the RDA Complex Citations Working Group (Research Data Alliance, 2024) has emerged with a focus on dataset collections in relation to research projects and publications, which aims to fulfill a specific use case that was identified in the planned work of the RDA Data Granularity WG. Finally, since research disciplines and data repositories are prioritizing differing and sometimes conflicting ways to describe granularity (e.g., structural components or dimensions), the WG suggests that disciplinary-focused entities build upon these recommendations by devising more specific standards or best practices for increased interoperability within a field, for applications such as federated systems and citation metrics.

Completing this work identified several potential areas of future work, including:

- Devise standard methods for evaluating data granularity approaches used (e.g., checklists, self-assessments, etc.).
- Further recommendations by other RDA groups specific to the granularity context within their focused scope.

In conclusion, given the complex and multi-leveled nature of research data, data granularity is a core issue that will benefit from these recommendations, as well as require continued attention and evolution in the years to come.

Acknowledgements

The RDA Data Granularity Working Group would like to express gratitude to all who assisted in its work leading to these outputs, including:

- The groups doing the foundational work that led to the WG: the RDA Data Discovery Paradigms Interest Group, including its Task Force on data granularity, and Earth Science Information Partners (ESIP).
- All members of the Data Granularity WG, past and present.
- The range of RDA Groups, as well as external entities, whose work and outputs were a core contributor to the WG's work (see Appendix 1).
- The support provided by the RDA community and structures.
- re3data, which was instrumental to the success of the survey. The WG is indebted to them for their contributions, and take full responsibility for any errors that may have occurred.
- Various members of the community who may have contributed in other ways: attendees of RDA plenary sessions, survey respondents, and those who informed the report and its use cases through informal conversations.

Funding:

- This work was developed as part of the RDA TIGER 'Research Data Alliance facilitation of Targeted International working Groups for EOSC-related Research solutions', funded by HORIZON-INFRA-2022-EOSC-0 Grant Agreement 101094406, and the authors acknowledge the support provided within the project partners and resources.
- Supported by the projects KonsortSWD and Base4NFDI which are funded by the German Research Foundation (DFG) within the framework of the NFDI – project numbers: 442494171, 521453681, 521460392, 521462155, 521463400, 521466146, 521471126, 521473512, 521474032, 521475185, 5214576232

Appendices

Appendix 1: Documents from Related RDA Groups

Given the integral, cross-cutting nature of granularity across a range of issues, the WG leveraged and built upon existing RDA work amongst a range of groups (see below). This list is but a selection of the key related groups drawn upon (data granularity potentially could arise across all topics). Other documents were reviewed (including those listed in the original case statement); following is the selected list specifically drawn upon for these outputs.

RDA Group	Relationship to Granularity	Key Resource(s)
Data Citation WG	Identifying particular segments of data	Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) https://doi.org/10.15497/RDA00016 (Rauber et al., 2018)
Data Discovery Paradigms IG	Data granularity is an important aspect of required metadata to support user requirements for data discovery	Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories http://doi.org/10.5334/dsj-2019-003 RDA IG Data Discovery Paradigms IG: Use Cases Data https://doi.org/10.5281/zenodo.1050976 (de Waard et al., 2017; Wu et al., 2019)
Data Foundation and Terminology WG	Defining different types (and granularities) of data objects	Data Foundation and Terminology Work Group Products https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF (Berg-Cross et al., 2019)
Data Type Registries WG	Data types at all levels of granularity	Data Type Registries Working Group Output https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458 (Lannom et al., 2018)

RDA Group	Relationship to Granularity	Key Resource(s)
Data Usage Metrics WG	Reporting data usage at different levels of granularity	Code of Practice for Research Data Usage Metrics, Release 1 https://doi.org/10.7287/peerj.preprints.26505v1 (Fenner et al., 2018)
Data Versioning WG/IG	Identification and versioning of different granularities of data, including use cases	Versioning Data is About More than Revisions: a Conceptual Framework and Proposed Principles http://doi.org/10.5334/dsj-2021-012 Compilation of Data Versioning Use Cases from the RDA Data Versioning Working Group, Version 1.1 https://www.rd-alliance.org/group_output/compilation-of-data-versioning-use-cases-from-the-rda-data-versioning-working-group/ Principles and Best Practices in Data Versioning for All Data Sets Big And Small, Version 1.1 https://www.rd-alliance.org/group_output/principles-and-best-practices-in-data-versioning-for-all-data-sets-big-and-small/ (Klump et al., 2020a, 2020b, 2021)
RDA/TDWG Metadata Standards for Attribution of Physical and Digital Collections Stewardship	Aggregating metrics for research products	RDA/TDWG Attribution Metadata Working Group: Final Recommendations https://www.rd-alliance.org/group_output/rda-tdwg-attribution-metadata-working-group-final-recommendations/ (Thessen et al., 2019)

RDA Group	Relationship to Granularity	Key Resource(s)
RDA/WDS Publishing Data Services WG	Evolving relationships amongst digital objects	ICSU-WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations https://doi.org/10.15497/RDA00002 (Burton & Koers, 2018)
GO FAIR Discovery Implementation Group	Use cases for data discovery, many related to granularity	Stocktaking GO FAIR Discovery IN - Use Cases, Infrastructure. https://doi.org/10.5281/zenodo.5006525 (Mathiak et al., 2021)

Appendix 2: Schemas and Registries Reviewed

Many of these were discovered using the RDA Metadata Standards Catalog: https://rdamsc.bath.ac.uk/search (Metadata Standards Catalog, 2020).

Schema	Schema Reference
DataCite 4.4	https://schema.datacite.org/meta/kernel-4.4/ (DataCite Metadata Working Group, 2021)
ISO 19115: 2014	https://www.iso.org/standard/53798.html (Committee ISO/TC 211, 2014)
IGSN	http://schema.igsn.org/description/ (IGSN e.V., 2022)
Dublin Core	https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (Dublin Core, 2020)
re3data	http://doi.org/10.48440/re3.010 (Strecker et al., 2021)
EML 2.2.0	https://eml.ecoinformatics.org/ (Jones et al., 2019)
INSPIRE	https://knowledge-base.inspire.ec.europa.eu/index_en (European Commission, 2024)

Appendix 3: Use Case Spreadsheet Documentation

The accompanying use cases spreadsheet contains instructions for its use; a list of use cases structured according to the user story format, which themselves are linked to narrative scenario descriptions; definitions of concepts used; and instructions used by use case authors. For the most part, use case information consisted of a pairing of 1) a narrative scenario and 2) discrete user stories pulled from each narrative (e.g., generally 2-10 user stories per narrative).

The user stories were written in a structured format, a sentence capturing a specific need of an end user, written in the following structure (spanning across several columns):
--As a [role/who], I want to [goal/what], so that [benefit to be achieved/why].

The user stories were coded according to three key characteristics: roles, levels of granularity, and categories (related to type of use, roughly analogous to stages of the data lifecycle). For consistency and ease of analysis, the following controlled vocabularies were created for each of these characteristics, with the ability of the use case author to indicate and describe an “other” as needed. See Use Cases spreadsheet for definitions and further details for the following.

Roles

- Data producer
- Data consumer
- Data subject
- Repository
- Data manager/steward/curator
- Publisher
- Funder
- Government/Policy Maker
- Research administrator
- Most/all

Levels of Granularity

- Measurement
- Variable
- File
- Instrument (of data collection)
- Custom subset
- Dataset
- Custom combination
- Collection (of datasets)
- Most/all

Category/Type of Use

- Findability/Discovery
- Access

- Interoperability
- Analysis
- Identification/Citation/Metrics
- Curation/Deposit/Preservation
- Most/all

This information, as well as places to note additional information and challenges for each scenario, is presented in additional columns in the spreadsheet

As the use cases were written, the structure of the spreadsheet evolved over time, reflecting debates in many meetings about the best approach among members of the subgroup (given the multiple possible approaches). One prominent debate regarded what is the optimum level of granularity for the user stories themselves (on the spectrum from high-level user story sentences down to minute detail). In the end, authors aimed for the “middle of the road,” “modest” level of granularity for the user story format, in order to balance internal consistency, ambiguity, and other factors.

Appendix 4: Repository Survey Invitation

Survey on “Data Granularity” launched

Together with the RDA Working Group on “Data Granularity” [1], we are conducting this survey to explore how repositories and data services are dealing with data granularity and what types of data they store. This will enable us to better describe the size and scope of repositories as well as the heterogeneous types of data granularity policies. Please help us by filling out the survey: https://forms.gle/iSrcYhW51BgWD8wt7 until the 24th of April. We thank you for your time; it will take 10 minutes at most.

[1] https://www.rd-alliance.org/groups/data-granularity-wg

Appendix 5: Repository Survey Instrument

Survey for Data Granularity

Thank you for participating in this survey!
Privacy notice: We will ask questions about a service (an archive, database, portal, repository or other type of institution) you may be affiliated with. When publishing the results of the survey later, we will anonymize the service you are answering for. However, knowing the identity of the service will allow us to add public metadata to your answers, making this survey much shorter than it would be otherwise, and allowing us to identify duplicates.

* Indicates required question

1 - Do you work for or are in other ways affiliated with a service (an archive, database, portal, repository or other type of institution) that stores research data or metadata to research data in the broadest sense?

Mark only one oval.

Yes

No *Skip to question 40*

2 - Data service types

For brevity's sake, we will refer to the archive or other type of institution that stores data as a service.

2.1 - If your data service is registered at re3data.org, please give us the URL, e.g. https://www.re3data.org/repository/r3d100011062 , alternatively, please give us the homepage of your service.

If your service is indexed in re3data, please add the URL to the entry here. You can check if your service is indexed in re3data via the re3data search. (https://www.re3data.org/search). Privacy notice: When publishing the results of the survey later, we will anonymize the service you are answering for. However, knowing the identity of the service will allow us to add public metadata to your answers, making this survey much shorter than it would be otherwise, and allowing us to identify duplicates.

Would you like to have your data service registered at Re3data.org? Check in the link

how to join the largest global registry of research data repositories....
https://www.re3data.org/

2.2 - What is the data service's regional scope in terms of the data you offer? * Regional scope: Global, local e.g. country, institutional, project/instrument-related.

Check all that apply.

- Global
- Local: country, region
- Institutional
- Focus on one or more instrument(s)
- Focus on one or more project(s)
- Not applicable

Other:

2.3 - What is the disciplinary scope of your data service?

Check all that apply.

- All disciplines
- Two or more disciplines (e.g. biology and medicine)
- One discipline
- Special field within one discipline (e.g. only trees, only DNA)
- Other
- Not applicable

2.4 - Disciplinary scope details

Please detail the data service's disciplinary scope.

2.5 - How would you judge the size of your data collection compared to other services with a similar scope?

Mark only one oval.

We are the only service with this scope.

We are the largest service with this scope.

We are among the largest services with this scope.

There are comparable services, which are much larger than what we have.

Other:

3 - Entity type 1

Here we are interested in the digital objects and data products you handle in your service and on what level you make them available to your users. We are mostly interested in the structure these entities have, what you store in them and how they interrelate.

3.1 - What is an entity type that you have in your data service, e.g. research data, documents, measurements, code? Please use your own words if needed. You will have the opportunity to add more entity types later.

If you store more than one entity type, please add them in the

subsequent questions. *Mark only one oval.*

Datasets

Publications

Software

Collections

Domain-specific data, e.g. DNA

Database

Other:

3.2 - Do you have PIDs for this entity type? Please check all that apply.

Check all that apply.

AN - Accession number

ARK - Archival Resource Key

DOI - Digital Object Identifier

Github gist

GRID

Handle

IGSN - International Geo Sample Number

ISBN - International Standard Book Number

ISNI - International Standard Name Identifiers

ISNI-IA ISNI - International Authority

ISSN - International Standard Serial Number

PURL - Persistent Uniform Resource Locators

RAiD - Research Activity identifier

RRID - Research resource identifiers

SHA-1 hash - Secure Hash Algorithm 1

URL - Uniform Resource Locator

URI - Uniform Resource Identifier

URN - Uniform Resource Name

UUID - Universally Unique Identifiers

We have an internal ID

Other:

3.3 - Which granularity level would you say the entity represent? You may check more than one box.

Check all that apply.

One single measurement or result, e.g. a data point or single number

A series of single measurements or results, e.g. a row or column in a spreadsheet

One file with multiple measurements or results, e.g. a spreadsheet

A package of multiple files, which represent the results of one group or one source

A collection of results from different sources, which are bundled thematically, e.g. a study series running over many years

A collection of results from different sources, which are bundled for convenience It also includes documentation in written form, e.g. as PDF

It also includes code, e.g. for further analysis or to document data cleaning

Other:

3.4 - How many units of this entity type are in your data service?

If you use more then one way to count this entity type, please list all counts you have available. If you don't have the numbers available, please give an estimate.

3.5 - Do you have more than one entity type in your data service?

Mark only one oval.

Yes

No *Skip to question 36*

4 - Entity type 2

4.1 - What would you call this entity type? Please use your own words if needed. If you store more than one entity type, please add them in the subsequent questions.

Mark only one oval.

Datasets

Publications

Software

Collections

Domain-specific data, e.g. DNA

Other:

4.2 - Do you have PIDs for this entity type?

Check all that apply.

AN - Accession number

ARK - Archival Resource Key

DOI - Digital Object Identifier

Github gist

GRID

Handle

IGSN - International Geo Sample Number

ISBN - International Standard Book Number

ISNI - International Standard Name Identifiers

ISNI-IA ISNI - International Authority

ISSN - International Standard Serial Number

PURL - Persistent Uniform Resource Locators

RAiD - Research Activity identifier

RRID - Research resource identifiers

SHA-1 hash - Secure Hash Algorithm 1

URL - Uniform Resource Locator

URI - Uniform Resource Identifier

URN - Uniform Resource Name

UUID - Universally Unique Identifiers

We have an internal ID

Other:

4.3 - Which granularity level would you say these represent? You may check more than one box.

Check all that apply.

One single measurement or result, e.g. a data point or single number

A series of single measurements or results, e.g. a row or column in a spreadsheet

One file with multiple measurements or results, e.g. a spreadsheet

A package of multiple files, which represent the results of one group or one source

A collection of results from different sources, which are bundled thematically, e.g. a study series running over many years

A collection of results from different sources, which are bundled for convenience

It also includes documentation in written form, e.g. as PDF

It also includes code, e.g. for further analysis or to document data cleaning

Other:

4.4 - Is this entity type connected to the entity type you mentioned earlier?

Check all that apply.

Yes, it is a subset of the first type, e.g. the first type is a collection and the second type belongs to this collection

Yes, it represents a collection of entities of the first type, or vice versa the answer above. Yes, it is linked to the first type. We make these links explicit in the metadata.

Yes, it is linked to the first type implicitly.

No, the entity types are not linked to each other.

Other:

4.5 - How many units of this entity type are in your data service?

In what unit(s) do you describe the size of your data service, e. g. dataset, collection, etc.? Does your data service use more than one unit to describe its size? Please use your own words to specify these units.

4.6 - Do you have more than two entity type in your data service? *Mark only one oval.*

Yes

No *Skip to question 36*

5 - Entity type 3

5.1 - Please give us the name of an entity type you have not mentioned before. If you store more than one entity type, please add them in the subsequent questions.

Mark only one oval.

Datasets

Publications

Software

Collections

Domain-specific data, e.g. DNA

Other:

5.2 - Do you have PIDs for this entity type?

Check all that apply.

AN - Accession number
ARK - Archival Resource Key
DOI - Digital Object Identifier
Github gist
GRID
Handle
IGSN - International Geo Sample Number
ISBN - International Standard Book Number
ISNI - International Standard Name Identifiers
ISNI-IA ISNI - International Authority
ISSN - International Standard Serial Number
PURL - Persistent Uniform Resource Locators
RAiD - Research Activity identifier
RRID - Research resource identifiers
SHA-1 hash - Secure Hash Algorithm 1
URL - Uniform Resource Locator
URI - Uniform Resource Identifier
URN - Uniform Resource Name
UUID - Universally Unique Identifiers
We have an internal ID

Other:

5.3 - Which granularity level would you say these represent? You may check more than one box.

Check all that apply.

One single measurement or result, e.g. a data point or single number
A series of single measurements or results, e.g. a row or column in a spreadsheet
One file with multiple measurements or results, e.g. a spreadsheet
A package of multiple files, which represent the results of one group or one source
A collection of results from different sources, which are bundled thematically, e.g. a study series running over many years

A collection of results from different sources, which are bundled for convenience It also includes documentation in written form, e.g. as PDF

It also includes code, e.g. for further analysis or to document data cleaning

Other:

5.4 - Is this entity type connected to the entity types you mentioned earlier? *Check all that apply.*

Yes, it is a subset of one of the earlier type, e.g. the earlier type is a collection and this type belongs to this collection

Yes, it represents a collection that may include another type

Yes, it is linked to another type. We make these links explicit in the metadata.

Yes, it is linked to another type implicitly.

No, the entity types are not linked to each other.

Other:

5.5 - How many units of this entity type are in your data service?

In what unit(s) do you describe the size of your data service, e. g. dataset, collection, etc.? Does your data service use more than one unit to describe its size? Please use your own words to specify these units.

5.6 - Do you have more than these entity type in your data service?

Mark only one oval.

Yes

No *Skip to question 36*

6 - Entity type 4

6.1 - Please give us the name of an entity type you have not mentioned before. If you store more than one entity type, please add them in the subsequent questions.

Mark only one oval.

Datasets

Publications

Software

Collections

Domain-specific data, e.g. DNA

Other:

6.2 - Do you have PIDs for this entity type?

Check all that apply.

AN - Accession number
ARK - Archival Resource Key
DOI - Digital Object Identifier
Github gist
GRID
Handle
IGSN - International Geo Sample Number
ISBN - International Standard Book Number
ISNI - International Standard Name Identifiers
ISNI-IA ISNI - International Authority
ISSN - International Standard Serial Number
PURL - Persistent Uniform Resource Locators
RAiD - Research Activity identifier
RRID - Research resource identifiers
SHA-1 hash - Secure Hash Algorithm 1
URL - Uniform Resource Locator
URI - Uniform Resource Identifier
URN - Uniform Resource Name
UUID - Universally Unique Identifiers
We have an internal ID

Other:

6.3 - Which granularity level would you say these represent? You may check more than one box.

Check all that apply.

One single measurement or result, e.g. a data point or single number

A series of single measurements or results, e.g. a row or column in a spreadsheet

One file with multiple measurements or results, e.g. a spreadsheet

A package of multiple files, which represent the results of one group or one source

A collection of results from different sources, which are bundled thematically, e.g. a study series running over many years

A collection of results from different sources, which are bundled for convenience It also includes documentation in written form, e.g. as PDF

It also includes code, e.g. for further analysis or to document data cleaning

Other:

6.4 - Is this entity type connected to the entity types you mentioned earlier? *Check all that apply.*

Yes, it is a subset of one of the earlier type, e.g. the earlier type is a collection and this type belongs to this collection

Yes, it is a collection that may include another type

Yes, it is linked to another type. We make these links explicit in the metadata.

Yes, it is linked to another type implicitly.

No, the entity types are not linked to each other.

Other:

6.5 - How many units of this entity type are in your data service?

In what unit(s) do you describe the size of your data service, e. g. dataset, collection, etc.? Does your data service use more than one unit to describe its size? Please use your own words to specify these units.

6.6 - Do you have more than these entity type in your data service? *Mark only one oval.*

Yes

No *Skip to question 36*

7 - Entity type 5

7.1 - Please give us the name of an entity type you have not mentioned before. If you store more than one entity type, please add them in the subsequent questions.

Mark only one oval.

Datasets

Publications

Software

Collections

Domain-specific data, e.g. DNA

Other:

7.2 - Do you have PIDs for this entity type?

Check all that apply.

AN - Accession number
ARK - Archival Resource Key
DOI - Digital Object Identifier
Github gist
GRID
Handle
IGSN - International Geo Sample Number
ISBN - International Standard Book Number
ISNI - International Standard Name Identifiers
ISNI-IA ISNI - International Authority
ISSN - International Standard Serial Number
PURL - Persistent Uniform Resource Locators
RAiD - Research Activity identifier
RRID - Research resource identifiers
SHA-1 hash - Secure Hash Algorithm 1
URL - Uniform Resource Locator
URI - Uniform Resource Identifier
URN - Uniform Resource Name
UUID - Universally Unique Identifiers
We have an internal ID

Other:

7.3 - Which granularity level would you say these represent? You may check more than

one box.

Check all that apply.

One single measurement or result, e.g. a data point or single number

A series of single measurements or results, e.g. a row or column in a spreadsheet

One file with multiple measurements or results, e.g. a spreadsheet

A package of multiple files, which represent the results of one group or one source

A collection of results from different sources, which are bundled thematically, e.g. a study series running over many years

A collection of results from different sources, which are bundled for convenience It also includes documentation in written form, e.g. as PDF

It also includes code, e.g. for further analysis or to document data cleaning

Other:

7.4 - Is this entity type connected to the entity types you mentioned earlier? *Check all that apply.*

Yes, it is a subset of one of the earlier type, e.g. the earlier type is a collection and this type belongs to this collection

Yes, it is a collection that may include another type

Yes, it is linked to another type. We make these links explicit in the metadata.

Yes, it is linked to another type implicitly.

No, the entity types are not linked to each other.

Other:

7.5 - How many units of this entity type are in your data service?

In what unit(s) do you describe the size of your data service, e. g. dataset, collection, etc.? Does your data service use more than one unit to describe its size? Please use your own words to specify these units.

7.6 - Is there anything you would want to add about the entity types you have in your data service?

8 - Issues with Granularity

8.1 - Who chooses the granularity level of the entities in your data service? Please check all that apply.

Check all that apply.

- The data service can choose the granularity level freely, repackaging files as needed
- We have guidelines for data providers, which are heavily enforced
- We have guidelines for data providers, but they are not always followed
- We expect data providers to adhere to the example set by others, but we do not enforce
- Data providers can do whatever they want
- Other:

8.2 - Would you like to have stricter rules for data granularity in your data service? *Mark only one oval.*

- The rules we have are already too strict
- No, the rules are just right
- We have some rules or informal guidelines, but they could be stricter
- We have some rules or informal guidelines, but they could be more heavily enforced
- Other:

8.3 - How would you rate the homogeneity of the granularity within and between the entities your data service provides?

Mark only one oval.

We mix all types of things together 1 2 3 4 5 All entities are exactly the same

8.4 - What are the challenges in your domain with regards to data granularity? *Check all that apply.*

- Discovery, e.g. results with low granularity make results with high granularity hard to find

Citation too specific, e.g. researcher would prefer to cite less entities

Citation too unspecific, e.g. researcher would prefer to cite more specific parts of the data

Reproducibility, e.g. researchers are unable to replicate results, because the data citations are unsuitable

Duplication, e.g. there are multiple entries for more or less the same data

Other:

9 - Thank you for your participation!

If you have questions about the survey, please either contact the co-chair of the RDA Working Group for Data Granularity Brigitte Mathiak (brigitte.mathiak@gesis.org) or Dorothea Strecker from re3Data (dorothea.strecker@hu-berlin.de).

9.1 - Please add any comments and thoughts you might have. The answers from this question will not be published. Likewise, indicate anything you might want us to keep private.

END OF APPENDICES

References

Albertoni, R., Browning, D., Cox, S., Gonzalez Beltran, A., Perego, A., & Winstanley, P. (2024). *Data Catalog Vocabulary (DCAT)—Version 3*. https://www.w3.org/TR/vocab-dcat-3/

Atlassian. (2024). *User stories with examples and a template*. https://www.atlassian.com/agile/project-management/user-stories

Australian Bureau of Statistics. (2023). *Statistical terms and concepts*. https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts

Ball, A., & Duke, M. (2015). *How to Cite Datasets and Link to Publications* (DCC How-to Guides). Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/cite-datasets

Berg-Cross, G., Ritz, R., & Wittenburg, P. (2019). *Data Foundation and Terminology Work Group Products*. Zenodo. https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF

Burton, A., & Koers, H. (2018). *ICSU-WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations* (1.0). Zenodo. https://doi.org/10.15497/RDA00002

Cambridge University Press. (2024). *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*. https://dictionary.cambridge.org/

CODATA, R. (2017). *International Research Data Management glossary (IRiDiuM)*. https://codata.org/initiatives/working-groups/standard-glossary-for-research-data-management-iridium/

Committee ISO/TC 211. (2014). *ISO 19115-1:2014—Geographic information—Metadata*. International Organization for Standardization. https://www.iso.org/standard/53798.html

DataCite. (2024). *DataCite Event Data* [Dataset]. https://support.datacite.org/docs/eventdata-guide

DataCite Metadata Working Group. (2021). *DataCite Metadata Schema for the Publication and*

Citation of Research Data and Other Research Outputs (4.4) [Dataset]. DataCite e.V. https://doi.org/10.14454/fxws-0523

Dataverse Users Community. (2022). *DOI Attribution to Datasets Vs Datasets and Individual Files* [Online post]. Dataverse Users Community. https://groups.google.com/g/dataverse-community/c/hFZRMXJ_-CE?pli=1

DDI Alliance. (2024). *DDI CDI: Cross-Domain Integration | Data Documentation Initiative*. https://ddialliance.org/Specification/ddi-cdi

de Waard, A., Khalsa, S. J., Psomopoulos, F., & Wu, M. (2017). *RDA IG Data Discovery Paradigms IG: Use Cases Data* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.1050976

Dublin Core. (2020, May 21). *DCMI: Dublin Core™*. https://dublincore.org/specifications/dublin-core/

Earth Science Data Systems, N. (2015, April 10). *EOSDIS Glossary | Earthdata* [Basic Page]. Earth Science Data Systems, NASA. https://www.earthdata.nasa.gov/learn/glossary

European Commission. (2024). *Infrastructure for Spatial Information in the European Community Metadata Standard* [Dataset]. https://knowledge-base.inspire.ec.europa.eu/index_en

eurostat. (2024). *Statistics Explained*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_Page

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P., & Chodacki, J. (2018). Code of Practice for Research Data Usage Metrics, Release 1. *PeerJ Preprints*, *6:e26505v1*. https://doi.org/10.7287/peerj.preprints.26505v1

GFZ German Research Centre For Geosciences, Humboldt-Universität Zu Berlin, Karlsruhe Institute Of Technology (KIT), Purdue University Libraries, Bertelmann, R., Buys, M., Cousijn, H., Dierolf, U., Elger, K., Fenner, M., Ferguson, L. M., Fritze, F., Fuchs, C., Goebelbecker, H.-J., Gundlach, J., Kindling, M., Kloska, G., Klump, J., Kramer, C., …

van de Sandt, S. (2024). *re3data: Registry of Research Data Repositories*. https://www.re3data.org/

Global Biodiversity Information Facility. (2024). *Global Biodiversity Information Facility*. http://www.gbif.org

Gregory, K., Khalsa, S. J., Michener, W. K., Psomopoulos, F. E., Waard, A. de, & Wu, M. (2018). Eleven quick tips for finding research data. *PLOS Computational Biology*, *14*(4), e1006038. https://doi.org/10.1371/journal.pcbi.1006038

ICPSR. (2024). *Glossary of Social Science Terms*. https://www.icpsr.umich.edu/web/ICPSR/cms/2042

IGSN e.V. (2022). *International Generic Sample Number (IGSN) Metadata Kernel Registration Schema* (1.0) [Dataset]. IGSN e.V. http://schema.igsn.org/

Jones, M., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Earl, S., & Chong, S. (2019). *Ecological Metadata Language (EML)* (2.2.0) [Dataset]. KNB Data Repository. https://doi.org/10.5063/F11834T2

Keet, C. M. (2010). A Top-Level Categorization of Types of Granularity. In J. Yao (Ed.), *Novel Developments in Granular Computing: Applications for Advanced Human Reasoning and Soft Computation* (pp. 92–130). IGI Global. https://doi.org/10.4018/978-1-60566-324-1.ch005

Klump, J., Wyborn, L., Downs, R. R., Asmi, A., Wu, M., Rider, G., & Martin, J. (2020a). *Compilation of Data Versioning Use Cases from the RDA Data Versioning Working Group* (1.1) [Dataset]. Research Data Alliance. https://www.rd-alliance.org/group_output/compilation-of-data-versioning-use-cases-from-the-rda-data-versioning-working-group

Klump, J., Wyborn, L., Downs, R. R., Asmi, A., Wu, M., Rider, G., & Martin, J. (2020b). *Principles and Best Practices in Data Versioning for all Data Sets Big and Small*. Research Data Alliance.

https://www.rd-alliance.org/group_output/principles-and-best-practices-in-data-versioning-for-all-data-sets-big-and-small/

Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., & Asmi, A. (2021). Versioning Data is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*. https://doi.org/10.5334/dsj-2021-012

Lannom, L., Broeder, D., & Manepalli, G. (2018). *RDA Data Type Registries Working Group Output*. Zenodo. https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

Mathiak, B., Juty, N., Heger, T., Di Donato, F., Jeschke, J., Widmann, H., Flügel, A., Culina, A., Bardi, A., & Kraker, P. (2021). *Stocktaking GO FAIR Discovery IN - Use cases, infrastructure* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.5006524

Metadata Standards Catalog. (2020). *Metadata Standards Catalog*. https://rdamsc.bath.ac.uk/

Murphy, & Parsons. (2021). *Finer Granularity Means Better Data: A Crowdsourcing Lab Experiment*. Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, Copenhagen, Denmark. https://ceur-ws.org/Vol-2932/paper5.pdf

National Information Standards Organization. (2024). *CRediT (Contributor Roles Taxonomy)*. CRediT. https://credit.niso.org/

Rauber, A., Asmi, A., van Uytvanck, D., & Proell, S. (2018, August). *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)*. Zenodo. https://doi.org/10.15497/RDA00016

Rauber, A., Asmi, A., van Uytvanck, D., & Pröll, S. (2020). Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of the IEEE Technical Committee on Digital Libraries*, *12*(1). https://doi.org/10.5281/zenodo.4048304

Rauber, A., Gößwein, B., Zwölf, C. M., Schubert, C., Wörister, F., Duncan, J., Flicker, K., Zettsu, K., Meixner, K., McIntosh, L. D., Pröll, S., Miksa, T., & Parsons, M. A. (2021). Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. *Harvard Data*

Science Review, *3*(4). https://doi.org/10.1162/99608f92.be565013

RDA Data Discovery Paradigms Granularity Task Force. (2019a). *Data Granularity Guidance—DRAFT* [Unpublished internal company document].

RDA Data Discovery Paradigms Granularity Task Force. (2019b). *Glossary/definitions* [Unpublished internal company document].

RDA Data Granularity Working Group: Use Cases Subgroup. (2024). *Use Case Concept Definitions* [Unpublished internal company document].

Research Data Alliance. (2021). *Data Granularity WG Case Statement rev-002*. https://www.rd-alliance.org/rationale/data-granularity-wg/rev-002/

Research Data Alliance. (2024). *Complex Citations Working Group Home*. https://www.rd-alliance.org/groups/complex-citations-working-group/

Stevenson, A. (2010). *Oxford Dictionary of English*. Oxford University Press. https://doi.org/10.1093/acref/9780199571123.001.0001

Strecker, D., Bertelmann, R., Cousijn, H., Elger, K., Ferguson, L. M., Fichtmüller, D., Goebelbecker, H.-J., Kindling, M., Kloska, G., Nguyen, T. B., Pampel, H., Petras, V., Schabinger, R., Schnepf, E., Semrau, A., Trofimenko, M., Ulrich, R., Upmeier, A., Vierkant, P., … Witt, M. (2021). *Metadata Schema for the Description of Research Data Repositories: Version 3.1*. GFZ Helmholtz-Zentrum Potsdam. https://doi.org/10.48440/re3.010

Thessen, A., Woodburn, M., & Koureas, D. (2019). *RDA/TDWG Attribution Metadata Working Group: Final Recommendations*. https://www.rd-alliance.org/group_output/rda-tdwg-attribution-metadata-working-group-final-recommendations/

United Nations. (2024). *Handbook on Management and Organization of National Statistical Systems: Glossary*. https://unstats.un.org/capacity-development/handbook/html/Handbook/Glossary/Glossar

y.htm

Witt, M., Cannon, M., Lister, A., Segundo, W., Shearer, K., Yamaji, K., & Group, R. D. A. D. R. A. W. (2024, May). *RDA Common Descriptive Attributes of Research Data Repositories* (1.0). Zenodo. https://doi.org/10.15497/RDA00103

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal*. https://doi.org/10.5334/dsj-2019-003