RDA Data Versioning Interest Group Charter

Name of Proposed Interest Group:

Data Versioning IG

Introduction (A brief articulation of what issues the IG will address, how this IG is aligned with the RDA mission, and how this IG would be a value-added contribution to the RDA community):

The demand for reproducibility of research results and re-using data is growing, therefore it will become increasingly important for a researcher to be able to cite the exact version of the dataset that was used to underpin their research publication. The capacity of computational hardware infrastructures have grown and this has encouraged the development of concatenated seamless data sets where users can use web services to select subsets based on spatial and time queries, or other data attributes. Further, the growth in computer power has meant that higher-level data products can be generated in really short time frames. This means that we need a systematic way to refer to the exact version of a data set or data product that was used to underpin the research findings or was used to generate higher-level data products, including who developed and also funded it.

Versioning procedures and best practices are well established for scientific software and can be used to enable reproducibility of scientific results. The codebase of very large software projects does bear some semblance to large dynamic datasets. Are these practices suitable for datasets or do we need different practices for data versioning? The need for unambiguous references to specific datasets was recognised by the RDA Working Group on Data Citation, whose final report recognises the need for systematic data versioning practices.

This gap was discussed at a BoF meeting held at the RDA Plenary in September 2016 in Denver, resulting in the formation in 2017 of an RDA Interest Group on Data Versioning. A review of the recommendations by this RDA Data Versioning IG concluded that systematic data versioning practices were not available. In 2018 the Working Group was formed and first met at P12 in Gaborone. Its focus was on assessing current practices and compiled 39 use cases of data versioning across 33 organisations globally. In January 2020, the WG produced a white paper documenting these use cases and recommended practices (Klump, et al, 2020a). The WG delineated 6 high-level principles, which provided a high-level framework for guiding the consistent practice of data versioning and can also serve as guidance for data centres or data providers when setting up their own data revision and version protocols and procedures (Klump, et al, 2020b & 2021). To further the adoption of the outcomes, the proposed new interest group plans to contribute the use cases and recommended data versioning practices to other groups in RDA, W3C, and other emerging activities in this field.

User scenario(s) or use case(s) the IG wishes to address (what triggered the desire for this IG in the first place):

A lack of accepted data versioning practices has been recognised in different fields where reproducibility of research is a concern, e.g. data citation, virtual research environments as well as identity and provenance to enabling giving credit to all individual contributors, This problem was addressed in the precursor to this new group, the RDA Data Versioning WG. The WG produced a consultative white paper that developed a framework for data versioning based around 6 principles (Revision, Release, Granularity, Manifestation, Provenance and Citation). This work was subsequently published in the Data Science Journal in March 2021[1].

However, the previous work by the WG only delineated the six data versioning principles. Feedback from the community is now asking for actionable guidelines and recommendations to assist people in applying the principles.

At the same time, questions are increasingly being raised relating to the identity, authority and ethics of re-publishing, mirroring and re-using data. Many of these questions relate to uncertainty surrounding identification and attribution of the data that has been re-published or mirrored, in particular when a re-publishing or mirroring process involves revision of data or metadata. Hence there is an opportunity within the new IG to show how the principles and recommendations on versioning can be applied to help resolve questions of attribution, authority, and ethics arising from data re-publication, mirroring and re-using.

Objectives (A specific set of focus areas for discussion, including use cases that pointed to the need for the IG in the first place. Articulate how this group is different from other current activities inside or outside of RDA.):

The objective of the RDA Data Versioning IG is:

1. Promote the adoption of the data versioning principles;
2. Develop a set of recommendations for applying the principles; and
3. Analyze how these principles and recommendations can be applied to questions of attribution, authority, and ethics arising from data publication, sharing and re-using.

The RDA Data Versioning IG builds on the RDA Data Versioning WG, which has established a distinct field of work not covered by other RDA groups, yet seeking synergies with existing groups.

[1] Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R.R. and Asmi, A., 2021. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*, 20(1), p.12. DOI: http://doi.org/10.5334/dsj-2021-012

Participation (Address which communities will be involved, what skills or knowledge should they have, and how will you engage these communities. Also address how this group proposes to coordinate its activity with relevant related groups.):

The IG plans to engage with the broad data community to seek the adoption of the data versioning principles, and work with the community to develop a common understanding of data versioning and standard practices. We have identified the following RDA groups to collaborate:

- Data Citation WG to include its outputs into the collection of use cases,
- Data Foundations and Terminology IG to contribute and adopt data terminology/vocabulary
- Data Granularity WG to align up versioning and granularity level for varied usage contexts
- Sharing Rewards and Credit (SHARC) IG to contribute use cases to the SHARC IG and seek their expertise on credit and ethics of re-publishing.
- Software Source Code IG to align data versioning concepts
- We will continue to identify and seek collaboration with other RDA IG or WG that may have a focus on any of the data versioning principles. For example, one of the data versioning principles is to record and communicate data provenance information. We notice that the Research Data Provenance IG has been withdrawn. However, there was a BoF session at the RDA Plenary 17 with the topic Aligning Provenance Approaches Across Domains. We will seek collaboration with this group as well should this be taken off as an IG/WG.

Outside of RDA, this IG will seek collaboration with communities such as the ESIP Research Object Citation Cluster and the W3C.

Outcomes (Discuss what the IG intends to accomplish. Include examples of WG topics or supporting IG-level outputs that might lead to WGs later on.):

The IG has three specific goals:

1) to promote the data versioning principles;
2) to develop recommendations for the adoption of the data versioning principles; and
3) to provide a forum for discussing issues related to data versioning, data re-publication, mirroring and re-using data.

Work on these topics may result in contributions to other RDA groups or starting a new IG/WG where there is no overlap with existing RDA groups.

Mechanism (Describe how often your group will meet and how will you maintain momentum between Plenaries.):

The IG plans to meet bi-monthly over a virtual conference tool, at a time to be determined by members. The IG will also communicate asynchronously through the RDA group mailing list and via shared online documents.

The group co-chairs may meet more often to discuss the group progress and strategy to engage members.

Timeline (Describe draft milestones and goals for the first 12 months):

Month	Activity
P17 (April 2021)	● Discuss the scope of the IG ● Continue to seek support from RDA members
May 2021	● Submit the case statement and commence the community review
Jun - Aug 2021	● Refine the IG scope ● Formation of task groups aligned with deliverables ● Recruit additional members
Aug - Dec 2021	● Progress task forces ● Update progress at RDA P18
Jan-Dec. 2022	● Work toward the recommendation ● Collect use cases on issues related to the identity, authority and ethics of re-publishing and mirroring of data

Potential Group Members (Include proposed chairs/initial leadership and all members who have expressed interest):

Membership of the Data Versioning IG will include the existing members of the Data Versioning WG.

Please note: If you are already a member of the Data Versioning WG you don’t need to add your name to the list below. We will populate it from the existing list of WG members.

FIRST NAME	LAST NAME	EMAIL	Affiliation	Country
Jens	Klump	jens.klump@csiro.au	Commonwealth Scientific and Industrial Research Organisation	Australia
Lesley	Wyborn	lesley.wyborn@anu.edu.au	National Computational Infrastructure Australian National University	Australia
Kirsten	Elger	kelger@gfz-postdam.de	GFZ German Research Centre for Geosciences	Germany
Mingfang	Wu	mingfang.wu@ardc.edu.au	Australian Research Data Commons	Australia
Robert	Downs	rdowns@ciesin.columbia.edu	CIESIN, Columbia University	USA
Hugh	Paterson III	i@hp3.me	Unaffiliated	USA/France