Data Versioning WG Case Statement

# Group name:

Data Versioning WG

# WG Charter

The demand for reproducibility of research results is growing, therefore it will become increasingly important for a researcher to be able to cite the exact version of the data set 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. 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 that was used to underpin the research findings, or was used to generate higher level products.

Versioning procedures and best practices are well established for scientific software and can be used 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 data sets 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](https://www.rd-alliance.org/groups/data-citation-wg.html), 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 of an Interest Group on data versioning. A review of the recommendations by the RDA Data Versioning IG (the precursor to this group) concluded that systematic data versioning practices are currently not available. The Working Group will produce a white paper documenting use cases and recommended practices, and make recommendations for the versioning of research data. To further adoption of the outcomes, the Working Group will contribute the use cases and recommended data versioning practices to other groups in RDA, W3C, and other emerging activities in this field. Furthermore, versioning concepts developed for research data will need to be brought in line with versioning concepts used in persistent identifier systems.

# Value Proposition

Data versioning is a fundamental element in work related to ensuring the reproducibility of research. Work in other RDA groups on data provenance and data citation, as well as the W3C Dataset Exchange Working Group, have highlighted that definitions of data versioning concepts and recommended practices are still missing. The outcomes of the Data Versioning Working Group will add a central element to the systematic management of research data at any scale by providing recommendations for standard practices in the versioning of research data. These practice guidelines will be illustrated by a collection of use cases.

# Engagement with existing work in the area

A lack of accepted data versioning practices has been recognised in different fields where reproducibility of research is a concern, e.g. data citation, data provenance, and virtual research environments. Versioning procedures and standard practices are well established for scientific software and can be used to facilitate the goals of reproducibility of scientific results. The Working Group will work with other groups within RDA and external on topics where data versioning is of importance to develop a common understanding of data versioning and standard practices.

Within RDA the Working Group will work together with the [Data Citation WG](https://www.rd-alliance.org/groups/data-citation-wg.html) to include its outputs into the collection of use cases, and with the [Data Foundations and Terminology IG](https://www.rd-alliance.org/groups/data-foundations-and-terminology-ig.html), the [Research Data Provenance IG](https://www.rd-alliance.org/groups/research-data-provenance.html), the [Provenance Patterns WG](https://www.rd-alliance.org/groups/provenance-patterns-wg), and the [Software Source Code IG](https://rd-alliance.org/groups/software-source-code-ig) to align data versioning concepts

The Working Group will work closely with the [W3C Dataset Exchange Working Group](https://www.w3.org/2017/dxwg/) to introduce the use cases collected by the RDA Data Versioning Working Group into the W3C Working Group’s [collection of use cases](https://www.w3.org/2017/dxwg/wiki/Use_Case_Working_Space) and [align versioning concepts](https://www.w3.org/2017/dxwg/wiki/General_versioning_considerations). Additionally, the RDA Versioning Working Group will work closely with the AGU FAIR Data Project, in particular Task Group E on Data Workflows.

# Work Plan

The outcome and deliverable of the Data Versioning WG will be a white paper documenting use cases, and recommending standard practices for data versioning. The use cases and recommendations will be aligned with the recommendations from other working groups in RDA, and external, where data versioning is of concern.

Milestones for the development of the document will be aligned with the coming RDA plenaries. The final document will be presented at the RDA Plenary in early 2019.

The Data Versioning WG will meet face-to-face at the RDA plenaries for broader discussions of the group’s findings and recommendations with other relevant RDA Groups. Between plenaries, the group will work online.

Besides sessions at the RDA plenaries, members of the working group will present the working group’s findings and recommendations at disciplinary conferences and in national working groups to achieve a broader community involvement in the development of the recommendations for data versioning.

The work on the data versioning white paper will be coordinated by the chairs of the working group. A collection of use cases will serve to illustrate the recommended practices for data versioning. The outcomes will be contributed as an addendum to the RDA Data Citation Recommendations to resolve differences between file-based and database-based applications.

Use cases collected by the Working Group will be fed into the W3C Dataset Exchange WG. This W3C Working Group has parallel timelines to the proposed RDA Data Versioning WG and will end in July 2019. It is now six months into its two year term.

# Adoption Plan

The Working Group will work with existing adopters to support the adoption process and document any successes, failures, and lessons learnt. The Working Group will collect feedback from adopters and make sure it is considered for inclusion in the outputs.

The Working Group will work closely with the [W3C Dataset Exchange Working Group](https://www.w3.org/2017/dxwg/) to introduce the use cases collected by the RDA Versioning WG into the W3C Working Group’s [collection of use cases](https://www.w3.org/2017/dxwg/wiki/Use_Case_Working_Space) and [align versioning concepts](https://www.w3.org/2017/dxwg/wiki/General_versioning_considerations). Initial outcomes will also be exchanged with the AGU FAIR Data Project.

# Initial Membership

The initial membership of the Data Versioning WG will be drawn from the membership of the Data Versioning IG. The initial membership will include links to other RDA groups, e.g. Research Data Provenance, Provenance Patterns WG, and Software Source Code IG (Mingfang Wu), to the W3C Dataset Exchange Working Group (Simon Cox), and the AGU FAIR Data Project (Jens Klump).

The Data Versioning WG will initially be led by Jens Klump (CSIRO), Lesley Wyborn (ANU), Robert Downs (Columbia University) and Ari Asmi (University of Helsinki).