Principles and best practices in data versioning for all datasets big and small

Version: 1.1

DOI: 10.15497/RDA00042

Authors: Jens Klump, Lesley Wyborn, Mingfang Wu, Robert Downs, Ari Asmi, Gerry Ryder, Julia Martin (Research Data Alliance Data Versioning Working Group)

Published: 06 April2020

Abstract: The demand for better reproducibility of research results is growing. With more data becoming available online, it will become increasingly important for a researcher to be able to cite the exact extract of the dataset that was used to underpin their research publication. However, while the means to identify datasets using persistent identifiers have been in place for more than a decade, systematic data versioning practices are currently not available. Without these, it is very hard for researchers to gain attribution and credit for their actual contributions to the collection, creation and publishing of individual datasets. Versioning procedures and best practices are well established for scientific software and can be used to enable reproducibility of scientific results.

The Research Data Alliance (RDA) Data Versioning Working Group produced this Final Report to document 39 use cases and current practices, and to make recommendations for the versioning of research data. To further adoption of the outcomes, the Data Versioning Working Group then contributed selected use cases and recommended data versioning practices to other groups in RDA and W3C. The outcomes of the RDA Data Versioning Working Group 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.

This revised report incorporates the feedback to the version 1.0 received from the RDA community review.

Language: English

Keywords: Data versioning, versioning procedures, use cases, data management

Citation and Download: Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder, G., Martin, J. (2020). Principles and best practices in data versioning for all data sets big and small. Version 1.1. Research Data Alliance. DOI: 10.15497/RDA00042.

Revised Report of the Research Data Alliance Data Versioning Working Group

Principles and best practices in data versioning for all datasets big and small

Version 1.1

Jens Klump, Lesley Wyborn, Mingfang Wu, Robert Downs, Ari Asmi, Gerry Ryder, Julia Martin

Version Information		
Release	Date	Description
Version 1.0	2020-01-16	Original submission Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder, G., & Martin, J. (2020). Principles and best practices in data versioning for all data sets big and small. Version 1.0. *Research Data Alliance*. DOI: 10.15497/RDA00042.
Community review: 2020-01-28 until 2020-02-28		
Version 1.1	2020-04-06	Updated with minor changes following community review **Recommended Citation:** Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder, G., & Martin, J. (2020). Principles and best practices in data versioning for all data sets big and small. Version 1.1. *Research Data Alliance*. DOI: 10.15497/RDA00042.

Executive Summary

The demand for better reproducibility of research results is growing. With more data becoming available online, it will become increasingly important for a researcher to be able to cite the exact extract of the dataset that was used to underpin their research publication. However, while the means to identify datasets using persistent identifiers have been in place for more than a decade, systematic data versioning practices are currently not available. Without these, it is very hard for researchers to gain attribution and credit for their actual contributions to the collection, creation and publishing of individual datasets. Versioning procedures and best practices are well established for scientific software and can be used to enable reproducibility of scientific results.

The Research Data Alliance (RDA) Data Versioning Working Group produced this Final Report to document 39 use cases and current practices, and to make recommendations for the versioning of research data. To further adoption of the outcomes, the Data Versioning Working Group then contributed selected use cases and recommended data versioning practices to other groups in RDA and W3C. This revised report incorporates the feedback received from the RDA community review.

From initial data acquisition, there can be multiple levels of processing applied and each level of processing can then have multiple versions created. This Final Report applies the Functional Requirements for Bibliographic Records (FRBR) to provide a conceptual framework with a set of data versioning principles developed around the FRBR concepts of the 'work', the 'expression', the 'manifestation' and the 'item'.

The two key recommendations for data versioning are:

1. Be clear about which dataset is to be identified and for what purpose; and
2. Communicate the significance of the change to the designated user community[1] of the dataset.

The versioning principles that emerged from the analysis of the 39 use cases are:

Revision (version control):

- A new instance of a dataset that is produced in the course of data production or data management that is different from its precursor is called a "revision".
- A dataset revision should be identified.

Release (data products):

[1] **Designated User Community:** An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time. (CCSDS, 2012 page 1–11)

- The release of a new version of a dataset should be accompanied by a description of the nature and the significance of the change.
- The significance of this change will depend on the intended use of the data by its designated user community.
- Each new release of a data product should have a new identifier.

Granularity (aggregates, composites, collections and time series):

- Data may be aggregated and combined into collections or timeseries.
- The collection should be identified and versioned, as should be each of its constituent datasets.
- Entire time series should be identified, as should be time-stamped revisions.

Manifestation (data formats and encodings):

- The same dataset may be expressed in different file formats or character encodings without differences in content. While these datasets will have different checksums, the work expressed in these datasets does not differ, they are manifestations of the same work.
- Manifestations of the same work should be individually identified and related to their parent work.

Provenance (derived products):

- The definition of revisions and releases signifies that a dataset has been derived from a precursor and is part of the description of its lineage, or provenance.
- Provenance can be more complex than following a linear path. Information accompanying a dataset release should therefore contain information on the provenance of a dataset.

Citation:

- Include information of the Release in the citation. DataCite recommends to use semantic versioning, issue a new identifier with major releases, use the “alternate identifier” and “related identifier” elements to identify releases and how they relate to other datasets, e.g. whether it was derived from a precursor.
- Updating the metadata does not create a new version, it only changes the catalogue entry.

1 Introduction

The demand for better reproducibility of research results is growing. More and more data is becoming available online. In some cases, the datasets have become so large that downloading the data is no longer feasible. Data can also be offered through web services and accessed on demand. This means that parts of the data are accessed at a remote source when needed. In this scenario, it will become increasingly important for a researcher to be able to cite the exact extract of the dataset that was used to underpin their research publication. However, while the means to identify datasets using persistent identifiers have been in place for more than a decade, systematic data versioning practices are currently not available.

Versioning procedures and best practices are well established for scientific software (e.g. Fitzpatrick et al, 2009; Preston-Werner, 2013). The related Wikipedia article gives an overview of software versioning practices (Wikipedia, 2019). The codebase of large software projects does bear some semblance to large dynamic datasets. Are therefore versioning practices for code also suitable for datasets or do we need a separate suite of practices for data versioning? How can we apply our knowledge of versioning code to improve data versioning practices? This Data Versioning Working Group investigated to which extent these practices can be used to enhance the reproducibility of scientific results (e.g. Bryan, 2018).

The Research Data Alliance (RDA) Data Versioning Working Group produced this Final Report to document use cases and practices, and to make recommendations for the versioning of research data. To further adoption of the outcomes, the Working Group contributed selected use cases and recommended data versioning practices to other groups in RDA and W3C. The outcomes of the RDA Data Versioning Working Group 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 are illustrated by a collection of use cases.

This revised report incorporates the feedback received from the RDA community review.

2 Related work in RDA and the W3C

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 (W3C DXWG, 2017), have highlighted that definitions of data versioning concepts and recommended practices are still missing. 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 its concepts could be applied to other use cases to facilitate the goals of reproducibility of scientific results. The Data Versioning Working Group worked with other groups within RDA and external on topics where data versioning is of importance towards developing a common understanding of data versioning and recommended practices.

Within RDA the Data Versioning Working Group worked with the Data Citation Working Group to include its outputs (Rauber et al., 2016) into the collection of use cases, and with the Data Foundations and Terminology Interest Group, the Use Cases Coordination Group, the Research Data Provenance Interest Group, the Provenance Patterns Working Group, and the Software Source Code Interest Group to align data versioning concepts. The Data Versioning Working Group also worked closely with the W3C DXWG to introduce selected use cases collected by the RDA Data Versioning Working Group into the W3C Working Group's collection of use cases and align versioning concepts. Additionally, the RDA Data Versioning Working Group worked closely with the AGU Enabling FAIR Data Project, in particular, Task Group E on Data Workflows.

An important driver to have a closer look at data versioning came from the work of the RDA Working Group on Data Citation, whose final report recognises the need for systematic data versioning practices (Rauber et al., 2016). The RDA Data Citation Working Group aimed to address issues of identifying and citing a subset of large and dynamic data collection. The RDA Data Citation Working Group recommends to version and timestamp any updates to data items and assign identifiers to timestamped queries which then allow the retrieval of specific data subsets at any given point in time. The recommendations given here are well suited for relational databases that are accessed using database queries but are not as well suited for file-based data. 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 Interest Group (the precursor to this group) concluded that systematic data versioning practices are currently not available.

3 The FRBR model applied to data versioning

We apply the Functional Requirements for Bibliographic Records (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998) as the conceptual framework to derive data version principles to be discussed in the sections below. The FRBR is a conceptual entity–relationship model developed by the International Federation of Library Associations and Institutions (IFLA), the model provides "a clearly defined, structured framework for relating the data that are recorded in

bibliographic records to the needs of the users of those records" (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998).

In the digital era, the FRBR model is proving ideal not just in helping to distinguish multiple derivatives (versions) of an original dataset, but in establishing transparent provenance chains of how a particular dataset evolved from the initial collection of the original data through to its publication, curation and archiving**,** and more importantly, being able to provide attribution and accreditation to those researchers, institutions, funders, etc that were involved in the creation and subsequent preservation of each version.

3.1 The FRBR model

As shown in Figure 1, the FRBR model has four top level entities: Work, Expression, Manifestation and Item:

1) **A Work** is an abstract entity, representing a distinct intellectual or artistic creation. When the modification of a work involves a significant degree of independent intellectual or artistic effort, the result is viewed as a new work; variant texts incorporating revisions or updates to an earlier text are viewed simply as expressions of the same work (e.g. Shakespeare's Hamlet);
2) An **Expression** is the intellectual or artistic realisation of a work, a realisation is in the form of a text, or particular notes, phrasing, etc., so any change in intellectual or artistic content constitutes a change in expression, for example, each edition, language or media of Shakespeare's Hamlet (a work) is considered as an expression;
3) A **Manifestation** is the physical embodiment of an expression of a work, because of its physical nature, any changes in physical form (e.g. typeface, medium, container) is considered a new manifestation; and
4) An **Item** is a concrete entity, representing a single exemplar of a manifestation and being used to track specific items. Note that an Item can be one or more than one physical object (e.g. a monograph issued as two separately bound volumes).

FRBR as a conceptual model may be subject to interpretation when implementing it, for example, when asking what kind of change results in a new expression or a new work. Nevertheless, the FRBR model (entities, entity attributes and relation attributes) has two prominent advantages:

1. It guides the organisation, aggregation and description of resources for serving four information seeking tasks by users: find (a work), identify (an expression), select (a manifestation), and obtain (an item). If a user sends a query with attribute of an entity (e.g. a work - find Romeo and Juliet materials, or a manifestation - recordings of Romeo and Juliet play), a system with implementation of the FRBR model would be able to aggregate up or drill down to a required granularity level to satisfy a search need; and

2. It helps to improve interoperability across catalogues under the same model.

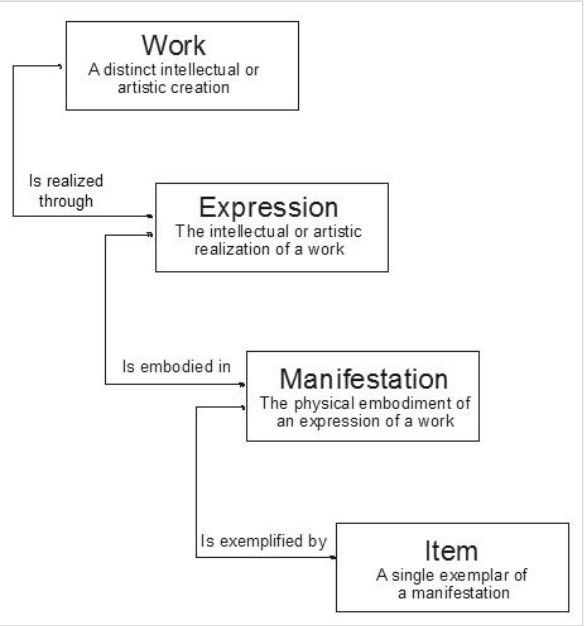

Figure 1. FRBR model: Relationship of Work, Expression, Manifestation and Item (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998)

3.2 Applying the FRBR model to scientific data

Our application of the FRBR model to data is based on earlier work by Hourclé (2009), who was one of the first to apply FRBR to scientific data, specifically to sensor-based scientific data. In his mapping, Hourclé suggests:

1) A Work is the calibrated state of a given observation;
2) An Expression is each resampling of an observation data (e.g. changes in resolution that alters the scaling of the values);
3) A Manifestation is a logical embodiment (instead of physical one in traditional library catalogue) of including aspects of how each individual datum is organised within a file or other package for distribution; and
4) An Item is defined by Hourclé as a logical item to be resolved such as URL, given that most data are digital objects (i.e. a specific physical manifestation).

Note that Hourclé's work focuses on the specific use case of remotely-sensed satellite data and the mapping may not be universally applicable to all data types because of the specific requirements of the use case. As an example, the determination of an element isotopic ratio on a mass-spectrometer has to go through several processing steps from the raw sensor output to a table to measurements and on to a higher aggregate product, but the types of corrections and conversions are completely different to those applied to satellite images. The concept of processing levels is still useful, but it has to be applied in a way that is specific to the use case.

We also disagree with Hourclé's application of the Work entity in the FRBR model to research data because in the FRBR definition the Work is an abstract entity. As a generalisation of Hourclé's work, we suggest an alternative mapping of the FRBR entities to data that takes into account concepts developed for the Observations and Measurements model (ISO 19156) (Cox, 2015) as follows:

1) A **Work** is the observation that results in the estimation of the value of a feature property, and involves application of a specified procedure, such as a sensor, instrument, algorithm or process chain;
2) An **Expression** of a work is the realisation of a work in the form of a logical data product. Any change in the data model or content constitutes a change in expression;
3) A **Manifestation** is the embodiment of an expression of a work, e.g. as a file in a specific structure and encoding. Any changes in its form (e.g. file structure, encoding) is considered a new manifestation; and
4) An **Item** is a concrete entity, representing a single exemplar of a manifestation, e.g. a specific data file in an individual, named data repository. An Item can be one or more than one object (e.g. a collection of files bundled in a container object).

3.3 Applying the FRBR model to data revision and versioning

In the #ASTER use case submitted by Wyborn (Klump et al, 2020), Wyborn applied the FRBR model to the data versions and revisions and versions for data products from Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER http://asterweb.jpl.nasa.gov), a Japanese imaging instrument on board the USA TERRA satellite. From the raw ASTER instrument data a series of derivative versions have been produced and can, to some extent, be aligned with the defined NASA processing levels that range from Level 0 (L0) to L4, with L0 products being the raw data at full instrument resolution, whilst at higher levels, the data are converted into more useful parameters and formats and released as additional versions (NASA, 2019). The initial reduction of the ASTER L0 'instrument' data to L1B/L2 'reflectance' products by Japan Space Systems (JSS) Ground Data Segment (GDS www.gds.aster.ersdac.or.jp) involves the correction for instrument, illumination, atmospheric and geometric effects.

In late 2009 an Australian initiative supported by multiple organisations, took the JSS L1B/L2 versions and created an L3 continental scale mosaic version. A series of product masks/thresholds were then applied to this L3 mosaic to generate a suite of seventeen L4 geoscience mineral maps that included 14 ASTER VNIR/SWIR Geoscience and three ASTER TIR products (Cudahy, 2012). Each of these seventeen mineral maps is available in: 1) Band Sequential (BSQ) image format that can be restretched/processed; 2) more GIS compatible products in GeoTIFF format; and 3) netCDF format to optimise use for HPC and scientific analysis (Cudahy, 2012). Variations of these L4 data products have been made accessible from

multiple organisational websites as either file downloads for local processing and/or in situ access as web services.

Considering the complexity of the ASTER use case, in particular the different formats of each data product, combined with the multiple sites a product is released from (with each with different access mechanism), the FRBR model was applied to help ensure **reproducibility** (knowing the source of any version that was used in any subsequent analysis), **provenance** (knowing the sequential history of any evolved data product) and **attribution** (knowing which organisation/individual had funded and/or produced and/or was sustaining the release of any version).

In detail, the various entities along the Full-path of ASTER data use are as follows:

1) **The Work**: the work is all observations taken by the ASTER sensor on board the Terra (EOS AM-1) satellite.

2) **The Expression:** The work is expressed as a sequence of four processing levels, with levels L0, L1A, L1B and L2 produced by JSS, and the Australian initiative producing the L3 Australian continental mosaic and from this mosaic, a set of seventeen mineral map L4 data products.

   In FRBR terminology, each level of processing (L0-L4), including each of the seventeen derived L4 mineral maps is considered to be an “expression” of the original ASTER “work”.

3) **The Manifestation**: Each of these seventeen L4 mineral maps is made available in three different formats that relate to different user requirements/ infrastructures/capabilities:
   a) Band sequential image (BSQ) files;
   b) GeoTIFF files; and
   c) Self-describing netCDF files for analysis at either continental scale and/or subsetting down to very small bounding boxes for local analysis.

   In FRBR terminology, each of these three formats is considered to be a separate “manifestation” of a particular “expression” of the “work”.

4) **The Item**: Items are generated from each manifestation and are delivered from a number organisational websites, e.g. CSIRO, Geoscience Australia, NCI, and individual State-Territory Geological Surveys (ref #ASTER for detail).

   In FRBR hierarchy terminology, each of these versions from each website is considered to be an “item” of each “manifestation” of each “expression”of the original “work”.

If each individual FRBR entity is given a globally unique PID, it will be possible to trace the Full-path of ASTER data for each individual item back to the original work (i.e. the ASTER mission), thus ensuring reproducibility and provenance tracking, but

more importantly, enabling attribution for and identity of any person/institution/ organisation involved in the development of any version.

4 Use cases

The Data Versioning Working Group collected 39 use cases from about 33 organisations and working groups that cover different research domains (e.g. social and economical science, earth science, environmental science, molecular bioscience) and different data types (Klump et al., 2020). The use cases describe current practices from data providers. These use case descriptions are useful in identifying differences in data versioning practices between data providers and highlighting encountered issues. We analysed the use cases in the context of the Data Versioning Working Group, but also registered them with the RDA Use Cases Group[2] for analysis that can be potentially carried by other Interest Groups/Working Groups with a different interest, e.g. data management analysis.

4.1 Identified issues

Through analysis of the use cases, we compiled the following list of issues or inconsistencies of practices across data producers:

- ***Issue 1:*** Although the definition of minor revision, substantial revision and major revision is context dependent, there should be a guideline on each. For example, a researcher who used an old version can be confident that a newer version with minor changes is not going to change his/her research outcome. (all use cases, e.g. #C-O-M, #C-F-H, #ASTER, #MT).

- ***Issue 2:*** There are different treatments if change in metadata results in a new version. (#BCO-DMO, #CSIRO, #USGS).

- ***Issue 3:*** Inconsistency in documenting version history: some comprehensive, some very light or not all. (all use cases, e.g. #DIACHRON, #VersOn).

- ***Issue 4:*** Inconsistency in naming and/or numbering each version: Data producers use various terms for version: e.g. Version 1,2; Collection 1,2; Release 1,2; Edition 1,2, vYYYYMMDD. What are the differences between each of these terms and what do these differences mean if they exist? (#NASA: EOSDIS and SEDAC #Molecular, #GA-EMC, #CMIP6, #RDA-DDC-R).

- ***Issue 5:*** The granularity of DOI: Should every revision have a DOI or a release/certain level of revision a DOI? (#USGS, #ESIP).

[2]RDA Use Cases Group: https://www.rd-alliance.org/groups/use-cases-group.html

- ***Issue 6:*** For a collection with multiple versions, a landing page may point to the latest version, all published versions, all published and archived versions. (#BCO-DMO, #NASA, #AAO, #GA-EMC, # Molecular).

- ***Issue 7:*** Format of citing versioned data: What version related information should be recommended for inclusion in a data citation (version number, data-access-URL, date-of-access, etc.)? (#IMOS).

4.2 Use Cases for W3C Dataset Exchange Working Groups (DXWG)

The W3C DXWG has documented lists of use cases including four use cases related to data versioning[3], namely: version definition, version identifier, version release date, and version delta; for the purpose of identifying current shortcomings and motivating the extension of the Data Catalog Vocabulary (DCAT) (Albertoni et al., 2019). To align with the W3C DXWG goal, we summarise six use cases that are related to metadata entity/attribute/vocabulary scope definition, as either an additional use case or more concrete requirements to existing W3C DXWG use cases. We discussed the following six use cases with the W3C DXWG for their consideration in their further iteration and prioritisation of use cases.

1. When changes are made to released data, data should be versioned. A previous version shouldn't be overwritten by the latest version, each version should be identifiable and retrievable. (#DIACHRON, #USGS #BCO-DMO).

2. When data has a new version, it should be easy for users to judge what kinds of changes have been made, so that users can 1) select the appropriate version, 2) assess if the changes would affect a research conclusion based on data from previous versions. (#DIACHRON, #USGS #BCO-DMO).

3. When a dataset is reprocessed using a different calibration, it should be possible for the user to identify datasets of different calibrations and retrieve a dataset of a specific calibration. (#AAO, #DEA).

4. The W3C working group recommends having a data revision when data are corrected, added (with or without changing data structure) and removed. (#USGS, #BCO-DMO, #CSIRO #Molecular).

   Consider also the following situations:
   a. new analytical and or processing techniques are applied to a select number of attributes/components of the existing dataset;
   b. models and derivative products are revised with new data;
   c. the data itself is revised as processing methods are improved;
   d. there is no change to data but rather the data structure, format or scheme; and

[3] https://www.w3.org/TR/dcat-ucr/#RVSDF

e. data is processed with a different calibration.

In each above use case, should a new version or a new dataset be recommended?

5. Is it possible to have a vocabulary of changes that result in a new version, so that a newer version can be annotated with a standardised vocabulary to describe what changes have been made?

6. There are several terms used for revision, e.g. Version, Collection, Release, Edition. Should all (or some of) these terminologies be unified under the same name “version”? (#NASA, #Molecular, #GA-EMC).

5 Versioning Principles

The recommendations given by the RDA Data Citation Working Group (Rauber et al., 2016) start with a key concept for data versioning: “Apply versioning to ensure earlier states of datasets can be retrieved” (R1 - Data Versioning). Fundamental to this recommendation is the requirement for unambiguous references to specific versions of data used to underpin research results. In this concept, any change to the data creates a new version of the dataset. A simple way to determine whether two datasets differ would be to calculate and compare a checksum (R6 - Result Set Verification). However, just knowing that the bit streams of two datasets differ does not give us other essential information that we might need to know.

The Data Versioning Working Group analysed the versioning use cases (Klump et al., 2020) outlined in this report and compared these practices to the recommendations of the RDA Data Citation Working Group. In addition to differences in the bitstream between two versions, we found a number of additional questions that versioning practices try to address:

- What constitutes a change in a dataset? (Revision) (Issue 1, 2);
- What are the magnitude and significance of the change? (Release) (Issue 1);
- Are the differences in the bitstream due to different representation forms? (Manifestation) (Issue 1);
- If the data are part of a collection and which elements of the collection have changed? (Granularity) (Issue 1, 5);
- How do two versions relate to each other? (Provenance) (Issue 3, 4, 6); and
- How can we express information on versioning when citing data? (Issue 4, 7).

Data versioning communicates not only that a dataset has been changed, but also refers to the significance and magnitude of change and other aspects. This finding corresponds to the work published by the W3C DXWG (Albertoni, et al., 2019).

Version control (Revision)

As noted above, the recommendations given by the RDA Data Citation Working Group already states that any change to a dataset creates a new version of the dataset that needs to be identified. This may also require the minting of a persistent identifier for this new version. This practice of fine-granular identification of versions is derived from version control commonly applied to the management of software code where every change to the code is identified as a separate version, often called a “revision” or “build” (Fitzpatrick et al, 2009). In the case of software versioning, the revision or build number can change far more frequently than the version number of a “released” version.

Data Production (Release)

The production of a dataset might require several iterations, resulting in several revisions, after which a new data product is released. From the perspective of reuse, it is important to understand this new version is compatible with other dependencies that may exist. Practices such as Semantic Versioning (Preston-Werner, 2013) propose best practices that communicate the significance of the change, the degree of compatibility, and development stage.

Objects and Collections (Granularity)

A collection of data may be the result of successively generated datasets. The full set of aggregated data (data collection) may be organised in a number of sub-collections to be served by a data repository or archive (Hourclé, 2009). 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 (Hourclé, 2009, Klump et al., 2016). The granularity should be determined by the use case to provide a way (or ways) of identifying parts and versions whenever the practical need arises (Paskin, 2003). Entire time series should be identified as collections (Klump et al., 2016), as should be time-stamped revisions, if the series is updated frequently (Rauber et al., 2016).

Formats (Manifestation)

The format of software code is tightly coupled to its syntax rules and compilation into executable software. Yet, the same software, e.g. a word processing package, maybe available for different platforms. Data, in the same way, may be manifested in different formats for use in different workflows while all are “manifestations” of the same “expression” of a “work”. Different “manifestations” may be identified separately in addition to identifying the “work”.

Derived Products (Provenance)

For scientific reproducibility, it is essential to know if a dataset was derived from a precursor and if yes, how these two objects relate to each other. Knowledge of the

history of a piece of information is known as “provenance”. Using provenance, it should be possible to understand how a piece of information has changed and whether it is fit for the intended purpose or whether the information should be trusted (Taylor et al., 2015).

6 Recommendations

The key learning from the working group’s activities was to distinguish between versioning based on changes in a dataset (data revisions) and communicating the significance of these changes (data release) as part of the data lifecycle. Thus, the two key concepts in data versioning are (1) to be clear about which dataset is to be identified and (2) what we want to communicate about it to its designated user community.

Version control and revisions

A new instance of a dataset that is produced in the course of data production or data management that is different from its precursor is called a **revision** and it should be separately identified. A common best practice in software development is to identify these new instances as revisions of their precursors and issue a revision number. The recommendations of the RDA Data Citation Working Group give guidance on the identification of a dataset revision with a combination of timestamp of a query and each change made to data (Rauber et al. 2016). If the repository uses the concept of revisions in version control, it should communicate the revision to the user through its relevant catalogue entry.

Identifiers for dataset revisions

Because the production of a revised dataset produces a new entity with a new identity, the data producer or the data repository should consider issuing a new identifier. Whether the revision is encoded in the dataset’s persistent identifier will depend on the policy of the data repository. If repositories decide to mint a global persistent identifier such as DOI for each revision, repositories should follow DataCite's recommendation not to use mnemonics in identifiers.

Identifying releases of data products

In some cases, the production of a dataset can be quite complex. The dataset may go through a number of revisions before it is considered to be “final”. The publication of such a “final” version of a dataset is called a “release”.

The release of a new version of a dataset should be accompanied by a description of the nature and the significance of the change. The significance of this change will depend on the intended use of the data by its designated user community. For instance, the release of a new version could signify changes in the data format and its compatibility with existing data processing pipelines, or significant changes to the

content of the dataset. Concepts such as Semantic Versioning (Preston-Werner, 2013) describe a commonly used practice to communicate the significance of a version change in a dataset release and have been widely adopted in software development.

Identification of data collections

Datasets may be aggregated into collections or timeseries. These collections can be seen as “works of works” (Hourclé, 2009), similar to a journal series. Following this practice, the collection (work of works) should be identified and versioned, and so should be each of its constituent datasets (works) (Klump et al., 2016).

Some data collections, such as time series data, are expected to change over time as new data are added. Here, the entire time series should be identified, as should be time-stamped revisions, if the series is updated frequently (Rauber, et al., 2016). As not all changes are due to the addition of data over time, but may also be the result of corrections, recalibrations, etc. it is also recommended to adopt a dataset release policy for time series data.

Identifying manifestations of datasets

The same dataset may be expressed in different file formats or character encodings without differences in content. While these datasets will have different checksums, the work expressed in these datasets does not differ, they are manifestations of the same work (Hourclé, 2009). From the perspective of content it might be sufficient to identify only the work, and not its manifestations, but there might be technical considerations such as machine actionability that merit a machine actionable identification of different manifestations of a work and and their instances as items through persistent identifiers (Razum et al., 2009).

Provenance of datasets

The definition of revisions and releases to describe that a dataset has been derived from a precursor helps to describe its lineage, or provenance. Semantic versioning, and related versioning schemes, encode in their release numbers information about a dataset and its precursors. Provenance, however, can be more complex than following a linear path. Information accompanying a dataset release should therefore contain information on the provenance of a dataset.

Requirements for Data Citation

The DataCite metadata kernel has an optional element (Element 15) to record the version of a dataset. DataCite recommends to use semantic versioning and furthermore recommends to issue a new identifier with major releases (DataCite Metadata Working Group, 2018). DataCite leaves it to the data stewards to determine what major or minor releases are. DataCite further recommends to use the alternate identifier (optional Element 11) and related identifier (optional Element

12) elements to identify releases and how they relate to other datasets, e.g. whether it was derived from a precursor. Note that this is the minimum required for data citation by DataCite; repositories may opt to offer a richer description of release history and provenance of a dataset through other channels.

Changes in the Metadata

Sometimes the metadata of a dataset is changed. This may be due to the correction of the metadata, metadata elements added, or any other reason. If these changes do not change the bitstream of a dataset manifestation, a change in the metadata does not constitute a new version. In short: a change in the catalogue entry does not constitute a new work.

7 Summary and future directions

The work of this group started out from the recommendations published by the RDA Dynamic Data Citation Working Group. In the course of the lifetime of this Working Group and its precursors we were able to document use cases of data versioning and from these use cases extract five principles for data versioning. Understanding data versioning will help us identify the data used in research and attribute them to the parties involved in their creation and curation.

The data versioning principles are designed to inform recommendations for best practices in data versioning. The vocabulary developed around data versioning allows us to describe and discuss issues related to data versioning with greater precision and clarity. The use cases collected and analysed by the Data Versioning Working Group already highlight the need for guiding principles and best practices of data aggregations and collections, and the re-publication and mirroring of data. The implementation of the data versioning principles raises new questions about the attribution of data sources, incentives for data publication, and reproducibility of research.

The Data Versioning Working Group, together with the RDA community, discussed at the RDA Virtual Plenary 15 where to take this work next. A way forward could be the development of implementation guidelines, possibly through an RDA Interest Group. This Interest Group could also coordinate future updates to the Data Versioning Principles and implementation guidelines. The current Data Versioning Working Group closed at the RDA Virtual Plenary 15.

Resources

This section provides links and further information about the resources and examples discussed in the article.

- Data Citation Working Group: https://www.rd-alliance.org/groups/data-citation-wg.html
- Data Fundations and Terminology Interest Group: https://www.rd-alliance.org/groups/data-foundations-and-terminology-ig.html
- Use Cases Coordination Group : https://www.rd-alliance.org/groups/use-cases-group.html
- Research Data Provenance Interest Group: https://www.rd-alliance.org/groups/research-data-provenance.html
- Provenance Patterns Working Group https://www.rd-alliance.org/groups/provenance-patterns-wg
- Software Source Code Interest Group: https://rd-alliance.org/groups/software-source-code-ig
- W3C Dataset Exchange Working Group: https://www.w3.org/2017/dxwg/
- W3C Data Exchange Working Group: Collection of use cases: https://www.w3.org/2017/dxwg/wiki/Use_Case_Working_Space
- W3C Data Exchange Working Group: Versioning concepts: https://www.w3.org/2017/dxwg/wiki/General_versioning_considerations
- NASA Earth Observing System Data and Information System (EOSDIS) data processing levels: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

Acknowledgements

The chairs of the RDA Data Versioning WG would like to thank all who contributed use cases to the WG and joined the discussions at the plenary sessions and along the way. We also thank the members of the community for their constructive comments during the community review process.

Use cases were contributed by Natalia Atkins (IMOS), Catherine Brady (ARDC), Jeff Christiansen (QCIF), Martin Capobianco, Andrew Marshall and Margie Smith (GA), Bob Downs (Columbia University), Kirsten Elger and Damian Ulbricht (GFZ Potsdam), Ben Evans, Nigel Rees, Kate Snow and Lesley Wyborn (NCI), Siddeswara Guru (TERN), Julia Hickie (NLA), Dominic Hogan (CSIRO), Leslie Hsu (USGS), Paul Jessop (International DOI Foundation), Dave Jones (StormCenter Communications Inc.), Danie Kinkaide (BCO-DMO), Heather Leasor (ADA, ANU), Benno Lee (Rensselaer Polytechnic Institute), Heather Leasor (ADA, ANU), Simon Oliver (Digital Earth Australia), Andreas Rauber (Vienna University of Technology), Simon O'Toole (AAO), Martin Schweitzer (BoM).

Special thanks go to the Australian Research Data Commons for their support.

We also like to thank our RDA Secretariat and TAB Liaisons, Stefanie Kethers and Tobias Weigel, for their guidance and support.

References

Albertoni, R., Browning, D., Cox, S. J. D., Gonzalez-Beltran, A., Perego, A., Winstanley, P., et al. (2019). Data Catalog Vocabulary (DCAT) - Version 2 (W3C Proposed Recommendation). Cambridge, MA: World Wide Web Consortium (W3C). Retrieved from https://www.w3.org/TR/vocab-dcat-2/ 25 March 2020.

Bryan, J. (2018). Excuse Me, Do You Have a Moment to Talk About Version Control? The American Statistician, 72(1), 20–27. Retrieved from https://doi.org/10.1080/00031305.2017.1399928 25 March 2020.

CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS). Magenta Book (Recommended Practice No. CCSDS 650.0-M-2). 135 pp. Greenbelt, MD: Consultative Committee for Space Data Systems. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf 25 March 2020.

Cox, S. J. D. (2015). Ontology for observations and sampling features, with alignments to existing models. Semantic Web Journal, 8(3), 453–570. Retrieved from http://www.semantic-web-journal.net/content/ontology-observations-and-sampling-features-alignments-existing-models-0

Cudahy, T. (2012). Satellite ASTER Geoscience Product Notes for Australia (No. EP125895) (p. 26). Canberra, Australia: Commonwealth Scientific and Industrial Research Organisation. Retrieved from https://doi.org/10.4225/08/584d948f9bbd1 25 March 2020.

DataCite Metadata Working Group. (2018). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data (Version 4.2). 69 pp. Hannover, Germany: DataCite e.V. Retrieved from https://doi.org/10.5438/bmjt-bx77 25 March 2020.

Fitzpatrick, B., Pilato, C. M., & Collins-Sussman, B. (2009). Version Control with Subversion. Sebastopol, CA: O'Reilly Media, Inc. Retrieved from http://svnbook.red-bean.com/ 25 March 2020.

Hourclé, J. A. (2009). FRBR applied to scientific data. Proceedings of the American Society for Information Science and Technology, 45(1), 1–4. Retrieved from https://doi.org/10.1002/meet.2008.14504503102 25 March 2020.

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). Functional Requirements for Bibliographic Records (IFLA Series on Bibliographic Control No. 19) (p. 142). Munich, Germany: International Federation of Library Associations and Institutions. Retrieved from http://www.ifla.org/publications/functional-requirements-for-bibliographic-records 25 March 2020.

Klump, J., Huber, R., & Diepenbroek, M. (2016). DOI for geoscience data - how early practices shape present perceptions. Earth Science Informatics, 9(1), 123–136. Retrieved from https://doi.org/10.1007/s12145-015-0231-5 25 March 2020.

Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder, G., & Martin, J. (2020). Compilation of Data Versioning Use cases from the RDA Data Versioning Working Group. Version 1.1. *Research Data Alliance*. DOI: 10.15497/RDA00041.

NASA (2019). Earth Observing System Data and Information System (EOSDIS) data processing levels. Retrieved from https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels 23 March 2019.

Paskin, N. (2003). On Making and Identifying a "Copy." D-Lib Magazine, 9(1). Retrieved from https://doi.org/10.1045/january2003-paskin 25 March 2020.

Preston-Werner, T. (2013). Semantic Versioning 2.0.0. Retrieved March 7, 2019, from https://semver.org/spec/v2.0.0.html (Original work published May 29, 2011)

Rauber, A., Asmi, A., van Uitvanck, D., & Pröll, S. (2016). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) (Technical Report). Denver, CO: Research Data Alliance. Retrieved from https://doi.org/10.15497/RDA00016 25 March 2020.

Razum, M., Schwichtenberg, F., Wagner, S., & Hoppe, M. (2009). eSciDoc Infrastructure: A Fedora-Based e-Research Framework. In Research and Advanced Technology for Digital Libraries (Vol. 5714, pp. 227–238). Heidelberg, Germany: Springer Verlag. Retrieved from http://dx.doi.org/10.1007/978-3-642-04346-8_23 25 March 2020.

Rees, N., Evans, B., Conway, D., Seillé, H., Goleby, B., Wyborn, L., (2019). Capturing (via automation) the Sequential Processing Levels along multiple Full-paths of Magnetotellurics Data Use. Extended Abstract, eResearch Australasia Conference, Brisbane–Australia, 21-25 October. Retrieved from https://conference.eresearch.edu.au/wp-content/uploads/2019/10/2019-eResearch-Rees_et_al.pdf 21 December 2019.

Taylor, K., Woodcock, R., Cuddy, S., Thew, P., & Lemon, D. (2015). A Provenance Maturity Model. In R. Denzer, R. M. Argent, G. Schimak, & J. Hřebíček (Eds.), Environmental Software Systems. Infrastructures, Services and Applications (Vol. 448, pp. 1–18). Cham, Switzerland: Springer International Publishing. Retrieved from http://doi.org/10.1007/978-3-319-15994-2_1 25 March 2020.

Wikipedia. (2019). Software Versioning. Wikipedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Software_versioning&oldid=886437916 March 11, 2019.

W3C Dataset Exchange Working Group (DXWG). (2017). Retrieved from https://www.w3.org/2017/dxwg/wiki/Main_Page March 20, 2019.