Versioning data is about more than revisions : A conceptual framework and proposed principles

Klump, Jens

2021

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 , vol. 20 , no. 1 , 12 . https://doi.org/10.5334/dsj-2021-012

http://hdl.handle.net/10138/340762
https://doi.org/10.5334/dsj-2021-012

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles

SPECIAL COLLECTION: RDA

RESEARCH PAPER

JENS KLUMP
JULIA MARTIN
LESLEY WYBORN
ROBERT R. DOWNS
MINGFANG WU
ARI ASMI

Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

A dataset, small or big, is often changed to correct errors, apply new algorithms, or add new data (e.g., as part of a time series), etc. In addition, datasets might be bundled into collections, distributed in different encodings or mirrored onto different platforms. All these differences between versions of datasets need to be understood by researchers who want to cite the exact version of the dataset that was used to underpin their research. Failing to do so reduces the reproducibility of research results. Ambiguous identification of datasets also impacts researchers and data centres who are unable to gain recognition and credit for their contributions to the collection, creation, curation and publication of individual datasets.

Although 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. In this work, we analysed 39 use cases and current practices of data versioning across 33 organisations. We noticed that the term 'version' was used in a very general sense, extending beyond the more common understanding of 'version' to refer primarily to revisions and replacements. Using concepts developed in software versioning and the Functional Requirements for Bibliographic Records (FRBR) as a conceptual framework, we developed six foundational principles for versioning of datasets: Revision, Release, Granularity, Manifestation, Provenance and Citation. These six principles provide 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.

CORRESPONDING AUTHOR:
Jens Klump
Mineral Resources Business Unit, Commonwealth Scientific and Industrial Research Organisation, Perth, Australia
jens.klump@csiro.au

KEYWORDS:
Data Versioning; File formats; Provenance; Citation; Reproducibility; Attribution

TO CITE THIS ARTICLE:
Klump, J, Wyborn, L, Wu, M, Martin, J, Downs, RR, and Asmi, A. 2021. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*, 20: 12, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-012*

INTRODUCTION

The demand for better reproducibility of research results is growing. A dataset, small or large, is often revised to correct errors, apply new algorithms, add new surveys, etc. A single data product can be released in different formats and by multiple providers: each of these can then be the source of additional derived data products, often by different authors than its precursor. For the reproducibility of research it is therefore important to know:

1. The source of each version that was used in any subsequent analysis and the sequential history of any evolved data product (provenance); and
2. Which organisation or individual produced and is sustaining the release of any version (attribution).

As more and more datasets are becoming available online, the versioning problem is becoming acute. In some cases, datasets have become so large that downloading the data as a single file is no longer feasible. Using web services, datasets can be accessed and subsetted at a remote source when needed, but the user often has no knowledge whether, and when, the dataset they accessed online has been changed or updated.

Errors in research data have the potential to cause significant damage when used to inform decision making. A high-profile case was the retraction of a multi-national study on a potential COVID-19 treatment that had to be withdrawn because it could not be reproduced (Mehra et al, 2020). Similarly, another published study, which had used data from the same source to investigate a potential COVID-19 treatment, also had to be retracted (Ledford and van Noorden, 2020). A lack of best practices for data versioning has been recognised as a contributing factor in what has been called a 'reproducibility crisis' (Peng, 2011). For research to be reproducible, it is essential for a researcher to be able to cite the exact extract of the dataset that was used to underpin their research publication (Allison et al, 2016).

Versioning procedures and best practices are well established for scientific software (e.g., Fitzpatrick et al, 2009; Preston-Werner, 2013), and the codebase of large software projects bears some semblance to large dynamic datasets. The related Wikipedia article gives an overview of software versioning practices (Software versioning, 2019). We investigated whether these concepts could be applied to data versioning to facilitate the goals of reproducibility of scientific results.

Whilst the means for identifying datasets by using persistent identifiers have been in place for more than a decade, community-agreed and systematic data versioning practices are currently not available. Confusion exists about the meaning of the term 'version' which is often used in a very general sense. Our collection of use cases showed the term 'version' referring to all kinds of alternative artefacts and the relationships between them (Klump et al, 2020a), extending beyond the more common understanding of 'version' to refer primarily to revisions and replacements (Software versioning, 2019).

The work presented in this paper was undertaken within the Research Data Alliance (RDA) Data Versioning Working Group (Klump et al, 2020a, 2020b) that worked with other RDA Groups, such as the Data Citation and Provenance Patterns Working Groups, the Data Foundations and Terminology, Research Data Provenance and Software Source Code Interest Groups, the Use Cases Coordination Group, as well as the Dataset Exchange Working Group of the W3C to develop a common understanding of data versioning and recommended practices.

The work of the RDA Data Versioning Working Group presented in this paper aimed to collect and document use cases and practices to make recommendations for the versioning of research data, and to investigate to which extent these practices can be used to enhance the reproducibility of scientific results (e.g., Bryan, 2018). The outcomes from this work add a central element to the systematic management of research data at any scale by providing a conceptual framework and six principles that can be used to guide the versioning of research data.

RELATED WORK

Versioning has been an important concept for tracking changes and identifying the state of a resource, especially for digital objects that are constantly going through revisions and changes.

For example, it is established practice in the software community to use versioning systems such as Concurrent Versioning Systems (CVS) (see e.g., Fitzpatrick et al, 2009) to keep track of changes in source code, and to name software versions and releases according to the semantic naming and numbering protocol (Software versioning, 2019).

Following the versioning practices in software development, the data community has recognised that, for reproducibility, it is important to understand that a dataset has been changed in the course of its life cycle. DataCite recommends that a data publication includes the version in its citation and that two versions of a data product cross-reference each other through the relation types 'HasVersion' and 'IsVersionOf' (DataCite Metadata Working Group, 2018).

Several international standards bodies include data versioning in their recommended practices. The W3C Dataset Exchange Working Group (Dataset Exchange Working Group, 2017) gives definitions of data versioning concepts. Among the use cases documented by the W3C working group are four use cases that focus on versioning, including version definition, version identifier, version release date, and version delta, identifying current shortcomings and motivating the extension of the Data Catalog Vocabulary (DCAT) (Albertoni et al, 2019). This work includes the provenance of versioning, as described in PROV-O (Lebo et al, 2013), and the provenance, authoring, and versioning (PAV) ontology (Ciccarese et al, 2013).

Many data centres now include data versioning as an important aspect of their data management practices (e.g., ESIP Data Preservation and Stewardship Committee, 2019), as can be seen in the many use cases collected by the RDA Data Versioning Working Group (Klump et al, 2020a).

The RDA Data Citation Working Group addressed the question of how to identify and cite a subset of a large and dynamic data collection. The recommendations given by the RDA Data Citation Working Group (Rauber et al, 2016) include data versioning as a key concept: '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. R6 – Result Set Verification offers a simple way to determine whether two datasets differ by calculating and comparing checksums.

However, through our work in the RDA Data Versioning Working Group we recognised that determining changes in the bitstream of a dataset is only one aspect of what is commonly called 'versioning'. Just knowing that the bitstreams of two datasets differ does not give us other essential information that we might need to know to determine the nature or significance of the change, or how different data files relate to each other.

The analysis of use cases and prior work showed that, although there are current practices, standards and tracking methods for data versioning, a high-level framework for guiding the consistent practice of data versioning is still lacking. The conceptual framework and data versioning principles proposed in this paper are intended to fill this gap.

USE CASES

The RDA Data Versioning Working Group collected 39 data versioning practice use cases from 33 organisations from around the world that cover different research domains, such as social and economic science, earth science, and molecular bioscience, and different data types (Klump et al, 2020a). The use cases describe current practices reported by data providers. These use case descriptions are useful in identifying differences in data versioning practices between data providers and highlighting encountered issues. The hashed names that appear in the list below point to the use cases collected for our analysis and cited in (Klump et al, 2020a).

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

- ***Issue 1:*** What constitutes a new release of a dataset and how should it be identified? Consider the following situations:
  - a. There is no change to data but rather the data structure, format or scheme;
  - b. Revisions are completed to correct identified corrigenda and errata;
  - c. New analytical and or processing methods are applied to a select number of attributes/components of the existing dataset;

d. Data is processed with a different calibration or parameterisation of the processing algorithm;

e. Models and derived products are revised with new or updated data; and

f. The data itself is revised as processing methods are improved, e.g. by a new algorithm.

In each above use case, we observe inconsistencies in the practices as to whether a new release or a new dataset should be recommended (#DIACHRON, #USGS #BCO-DMO, #CSIRO #Molecular, #AAO, #DEA).

- ***Issue 2:*** What is the significance of the change from one version to the next?

  Although the definition of minor revision, substantial revision and major revision is context dependent, there should be a guideline on each, including how to identify corrigenda and errata. For example, a researcher who used an old version should be able to determine from the documentation of the changes whether a newer version with minor changes could change the outcome of their research (all use cases, e.g., #C-O-M, #C-F-H, #ASTER, #MT).

- ***Issue 3:*** Do changes in the metadata change the version of the associated dataset?

  There are inconsistent practices for treating metadata revision and the implications for the data described by the metadata. For some use cases, a change in the metadata initiates a new version of the data and creation of a new DOI. Other use cases argue that if only the metadata is updated, neither a new version of the data nor a new DOI is created (compare #BCO-DMO, #CSIRO, #USGS).

- ***Issue 4:*** What needs to be included in a versioning history?

  Inconsistency in documenting version history: some comprehensive, some very light or not all. 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, and 2) assess if the changes would affect a research conclusion based on data from previous versions (all use cases, e.g., #DIACHRON, #VersOn, #USGS #BCO-DMO).

- ***Issue 5:*** How should a version be named or numbered?

  Inconsistency in naming 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 6:*** 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? (#USGS, #ESIP)

- ***Issue 7:*** Which version does the landing page of a dataset point to?

  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 8:*** What versioning information should be included in a data citation?

  Version related information should, at a minimum, include a version number, data-access-URL, date-of-access, or other identifying information.

REVISION VS. RELEASE

In our analysis of the use cases we noticed that the term version, revision, and release were used almost interchangeably even though the three terms can mean different things across the use cases. Following common practices in software development (Software versioning, 2019), we distinguish between tracking revisions of and changes made to a dataset, and the editorial process of identifying a particular version of a dataset as a release.

Here we define a revision as identifying that a change is made to the bitstream of a dataset, while the release of a dataset is an editorial process that designates a particular revision being

published for reuse. The description of a dataset release must explain the significance of the changes to its designated user community. Not all revisions lead to a new release.

THE FRBR MODEL AND ITS APPLICATION TO DIGITAL DATA VERSIONING

The identification of versions is not unique to data and software but is relevant for other information resources too. The International Federation of Library Associations and Institutions (IFLA) Study Group on the Functional Requirements for Bibliographic Records developed a general conceptual framework to describe how information resources relate to each other in their Functional Requirements for Bibliographic Records (FRBR) (Study Group on the Functional Requirements for Bibliographic Records, 1998). We identified FRBR as a potential conceptual framework to inform our data versioning principles which will be discussed in the sections below. The FRBR model is a conceptual entity–relationship model to provide 'a clearly defined, structured framework for relating the data that are recorded in bibliographic records to the needs of the users of those records' (Study Group on the Functional Requirements for Bibliographic Records, 1998). ***Figure 1*** shows the definition of four top level entities and their relation: Work, Expression, Manifestation and Item.

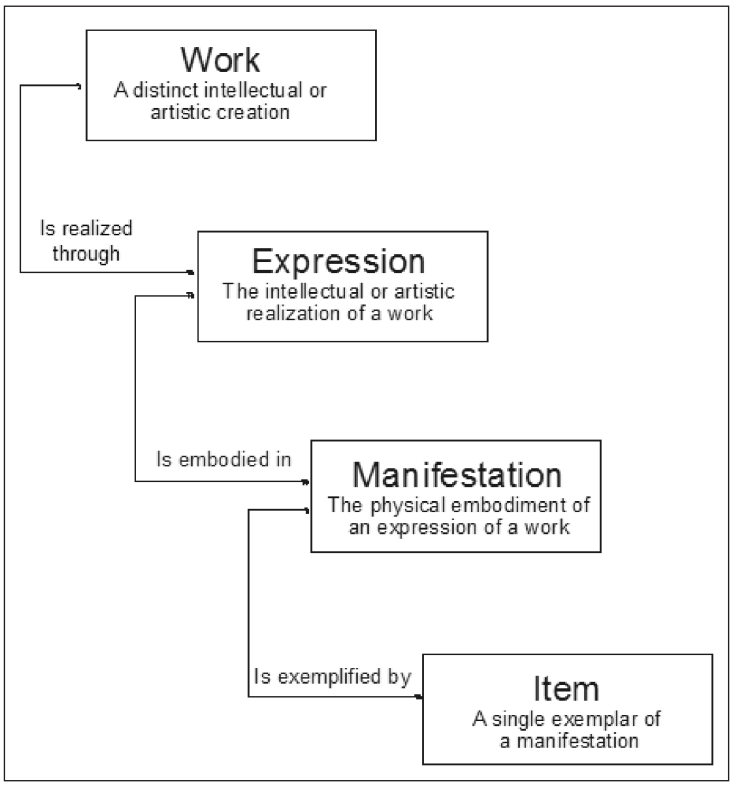

Figure 1 Relationship of Work, Expression, Manifestation and Item in the FRBR model (Study Group on the Functional Requirements for Bibliographic Records, 1998).

In the digital era, the FRBR model has the potential not only to distinguish multiple derivatives of an original dataset, but to also help establish transparent provenance chains describing how a particular dataset evolved from the initial collection of the original data through to its publication**,** and more importantly, to then be able to provide attribution and credit to those researchers, institutions, funders, etc who were involved in the creation of each individual version.

Hourclé (2009) was the first to apply the FRBR model to scientific data. Note that Hourclé's work focuses on the specific use case of remotely sensed satellite data and the mapping of processing levels (National Aeronautics and Space Administration, 2019) 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.

As a generalisation of Hourclé's work, we suggest an alternative mapping of the FRBR entities to data that also takes into account more recent concepts developed for the Observations and Measurements model (ISO 19156) (Haller et al, 2018).

1. A **Work** is the observation (e.g. an experiment) 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. In the FRBR model the work is an abstract entity;
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).

Taking the #ASTER use case as an example, Wyborn (in Klump et al, 2020a) applied the FRBR model combined with the well-established NASA processing levels (National Aeronautics and Space Administration, 2019) to document the Full Path of Data (Asch et al, 2018) used for the sequence of data products and data distributions derived from the original Japanese Space System (JSS) Advanced Spaceborne Thermal Emission and Reflectance Radiometer mission (ASTER – http://asterweb.jpl.nasa.gov, ***Figure 2***). In this use case, we use the term Full Path of Data to track how the dataset evolves starting with the capture of the original source data through the production of multiple derivative products and ultimately its distribution from multiple sources: each one of these 'products' will have its own data life cycle (Asch et al, 2018).

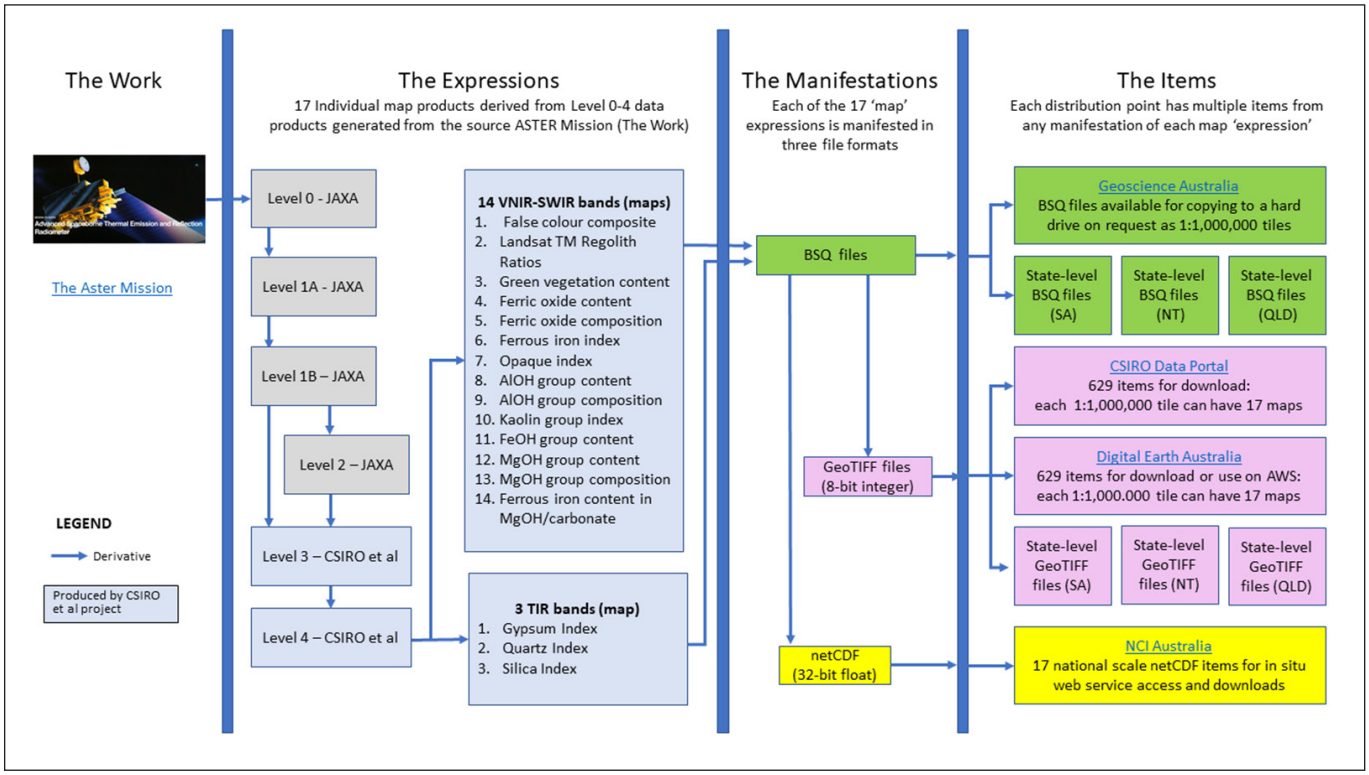

Figure 2 Schematic overview of the FRBR model applied to Full path of data use from the ASTER Mission. Each processing lavel (grey boxes) has its own data life cycle (e.g., (Wing, 2019)).

In late 2009, a national initiative supported by multiple Australian organisations produced a set of 17 ASTER National data products of the Earth's surface mineralogy of Australia that could be used from a continental scale (1:250,000) down to the scale of a mineral prospect (1:50,000). Each of these 17 mineral maps was available in 1) Band Sequential (BSQ) image format; 2) more GIS compatible products in GeoTIFF format; and 3) netCDF files to optimise use for analysis in High-Performance Computing (Cudahy, 2012).

Considering the complexity of the ASTER use case, in particular the different formats of each data product combined with the multiple sites each is released from, the FRBR model proved

to be useful. When combined with the use of unique persistent identifiers, the model can be used to help ensure both reproducibility by knowing the identity of each object, provenance by knowing the source of each version that was used in any subsequent analysis as well as the sequential history of any evolved data product, and attribution by knowing which organisation/individual had produced and was hosting the release of any version.

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

1. **The Work:** In this example, the Work is the observations taken by the ASTER sensor on board the Terra (EOS AM-1) satellite. This work is expressed in a number of data products.
2. **The Expression:** The reduction of the ASTER Level 0 (raw instrument data) to Level 1B or Level 2 products by JSS produced the 4 initial versions of the ASTER 'work', as shown in the grey boxes in ***Figure 2***. The Australian initiative then produced a set of mineral map data products from the Level 1B or Level 2 products (Cudahy, 2012). This involved applying a series of product masks/thresholds to generate a suite of geoscience mineral maps that included 14 ASTER VNIR/SWIR Geoscience products and three ASTER Thermal Infrared (TIR) products.

   In FRBR hierarchy terminology, each product at each processing level, such as L0 to L3, as well as each of 17 maps derived in L4 are considered to be an **Expression** of the **Work**. These represent 22 individual Expressions in total.
3. **The Manifestations:** Each of these 17 L4 mineral maps were made available in three different formats that relate to different user requirements/infrastructures/capabilities:

   i. Band sequential image (BSQ) files that can be restretched/processed;

   ii. GeoTIFF files that were generated by contrast stretching and colour rendering to national standards to generate more user-friendly GIS-compatible products; and

   iii. Self-describing netCDF files for analysis at full resolution and at continental scale: these could also be subsetted down to very small bounding boxes for local analysis at the prospect or local scale.

   In the terminology of the FRBR model, each set in each of the three formats is considered to be a **Manifestation** of each of the 17 L4 **Expressions** of the **Work**, resulting in 51 manifestations in total.
4. **The Items:** The file sizes of some of the national coverages were very large – the netCDF files are ~60 GB each. When these files were first created in 2012 they were too large to make available online as file downloads and a series of items, subsetting these files as standard 1:1,000,000 map tiles were generated from each manifestation and delivered from various organisational websites, e.g., CSIRO, Geoscience Australia, and State-Territory Geoscience Maps at the State level. The NCI instance still provided web service access to the large single files and the user could generate their own subset and either use it in situ or download it for processing it locally.

   Following the definition in the FRBR model, each file released from each location is considered to be an **Item** of a **Manifestation** of each of the 17 L4 **Expressions** of the original **Work**, represented by more than 1200 items in total.

To show the general nature of the applicability of FRBR to research data, we applied the pattern to a dataset from PANGAEA (König-Langlo and Gernandt, 2008). This dataset is a collection of 426 individual data sets of tabulated data documenting a meteorological radiosonde ascent launched from the Georg-Forster Antarctic Research Station operated by the German Democratic Republic. The data are stored in a relational database at the PANGAEA data repository. The tables for data delivery are generated on demand in a wide range of character encodings (Diepenbroek et al, 2002).

In the example of data in PANGAEA, the FRBR model can be applied as follows:

1. **The Work:** In this example, the Work is the observations taken during the radiosonde ascends. This work is expressed in a data product, i.e. (König-Langlo and Gernandt, 2008).
2. **The Expression:** The series of radiosonde ascends is expressed as a set of tabulated data.

3. **The Manifestations:** The data are offered by the PANGAEA database as a manifestation of the data in a specific character encoding.
4. **The Items:** The data Items are the bitstream of the data delivered by PANGAEA for download in the specified encoding.

THE SIX DATA VERSIONING PRINCIPLES

Using the FRBR model as our reference model, we analysed the issues as extracted from the use cases (Klump et al, 2020a), the work from W3C (Albertoni et al, 2019) and the RDA Data Citation Working Group (Rauber et al, 2016). While prior work focused on differences in the bitstream between versions, we found a number of additional questions that versioning practices try to address:

- What constitutes a change in a dataset? (Revision: Issues 1, 2, 3)
- What are the magnitude and significance of the change? (Release: Issues 1, 2)
- Are the differences in the bitstream due to different representation forms? (Manifestation: Issue 2)
- If the data are part of a collection and which elements of the collection have changed? (Granularity: Issues 2, 6)
- How do two versions relate to each other? (Provenance: Issues 4, 5, 7)
- How can we express information on versioning when citing data? (Citation: Issue 5, 8)

Note that we are specifically discussing changes in the dataset, not the metadata records in a catalogue that describe an individual dataset. Updating the metadata record does not create a new version in our model, it only changes the catalogue entry. Sometimes the metadata record of a dataset can be changed due to the correction of the metadata, metadata elements added, changing the location of the service endpoints or any other reason. If these changes do not change the bitstream of a dataset manifestation, a change in the metadata record does not constitute a new version.

The analysis of the use cases documented in (Klump et al, 2020a) demonstrated the need to distinguish between versioning based on changes in a dataset (data revisions) versus communicating the significance of these changes (data release) as part of an editorial process in the data lifecycle. In addition, we recognised that concepts from the FRBR model can be used to describe relationships between different versions of a dataset.

As an outcome from our analysis, we recommend the following 6 principles to address the identified issues in data versioning.

PRINCIPLE 1: VERSION CONTROL AND REVISIONS (REVISION)

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 (uniquely) identified. As noted in the discussion of prior work, 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 (Rauber et al, 2016). This may also require the minting of a persistent identifier for this new version.

This practice of fine-granular identification of revisions 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.

PRINCIPLE 2: IDENTIFYING RELEASES OF A DATA PRODUCT (RELEASE)

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 must be accompanied by a description of the nature and the significance of the change, along with a description of possible implications for use that could result from 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.

PRINCIPLE 3: 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 organised in a number of sub-collections to be served by a data repository or archive (Hourclé, 2009). The collection of works must be identified and versioned, and so shall be its constituent datasets or individual works (Klump et al, 2016).

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 is to 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).

PRINCIPLE 4: IDENTIFICATION OF MANIFESTATIONS OF DATASETS (MANIFESTATION)

The same dataset may be expressed in different file formats or character encodings, sometimes referred to as distributions (Albertoni et al, 2019), 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. From the perspective of content it might be sufficient to identify only the expressions of a 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 their instances as items through persistent identifiers (Razum et al, 2009).

PRINCIPLE 5: REQUIREMENTS FOR PROVENANCE OF DATASETS (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). Information accompanying a dataset release should therefore contain information on the provenance of a dataset.

PRINCIPLE 6: REQUIREMENTS FOR DATA CITATION (CITATION)

Data publications must include information about the Release in the citation and metadata. The DataCite metadata kernel (DataCite Metadata Working Group, 2018) has an optional element (Element 15) to record the version of a dataset. DataCite recommends using Semantic Versioning and furthermore recommends issuing a new identifier with major releases. DataCite leaves it to the data stewards to define major and minor releases. DataCite further recommends using 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; data centres and other repositories may opt to offer a richer description of release history and provenance of a dataset through other channels.

SUMMARY AND FUTURE DIRECTIONS

In this paper we present six principles on data versioning based on the analysis of 39 use cases (Klump et al, 2020a) and the FRBR framework (Study Group on the Functional Requirements for Bibliographic Records, 1998). The principles allow us to describe and discuss issues related to data versioning with greater precision and clarity, even when particular communities have different policies and procedures on data versioning. The six principles can be treated as a conceptual framework, yet each principle on its own is implementable as part of a data versioning policy or procedure.

Data publishers and providers must publish their data versioning policies and procedures to enable their users to identify the exact version of the data and any data extract that was used in a research project and in subsequent publications. Rigorous versioning procedures and policies will also enable proper attribution and credit to those parties involved in the creation, publication and curation of any data product and its precursors. Where data is accessible as online web services and/or are dynamically being constantly updated, as in a time series, it is essential that the data publisher/provider also makes available machine readable records of where they have made any changes to the dataset. This includes not only known changes to the data itself, but also changes in hardware and software, including versions of web service standards, that could also affect the data.

Versioning is also relevant to the application of the FAIR principles, particularly for the aspect of reuse (R1.2, Wilkinson et al, 2016). To interoperate or aggregate data sets from multiple sources, the exact version being merged needs to be known to ensure compatibility, reproducibility, provenance, and attribution. Precise and unambiguous versioning also facilitates the reuse of data, particularly if each version is clearly licensed and provides revision history and source information, as these in turn, help enable the user to define the quality of the version and whether the data are fit for their specific purpose. Similarly, adoption of such data versioning practices also can contribute to achieving transparency, responsibility, user focus, sustainability, and technology (TRUST), as described in the TRUST Principles for Digital Repositories (Lin et al, 2020).

We also expect to work with those implementing the data versioning principles in the future in a follow-up group to the RDA Data Versioning WG, and to verify and revise these principles, where necessary. In the next step, the data versioning principles should be developed into actionable recommendations by working with the community to develop domain specific policies on how to identify and communicate data versioning and by sharing examples of technical and policy implementations of the principles to develop the best practices for the versioning of datasets.

The analysis of the use cases and discussions within the community raised questions about the ethics of data duplication and re-publication, incentives for proper publication of all relevant data, particularly for the rawer precursor forms of derived data products, and above all, the reproducibility and replicability of research. This discussion highlights the need for documentation of best practices for the identification of data aggregations, data re-publication and mirroring of data to multiple sites. In future work we will explore issues related to defining the authoritative or canonical version of a dataset and correct attribution and citation of data sources.

ACKNOWLEDGEMENTS

This work was developed as part of the RDA Data Versioning Working Group. We acknowledge the support provided by the RDA community and structures and we would like to thank members of the group for their support for contributing the use cases and joining the discussions at the plenary sessions and along the way.

Use cases were contributed by:

Austria: Andreas Rauber (Vienna University of Technology)

Australia: Natalia Atkins (Integrated Marine Observing System, IMOS), Catherine Brady (Australian Research Data Commons, ARDC), Jeff Christiansen (Queensland Cyber Infrastructure Foundation, QCIF), Martin Capobianco, Andrew Marshall and Margie Smith (Geoscience Australia, GA), Ben Evans, Nigel Rees, Kate Snow and Lesley Wyborn (National Computational Infrastructure, NCI, Australian National University, ANU), Siddeswara Guru (Terrestrial Ecosystems Research Network, TERN),

Julia Hickie (National Library of Australia, NLA), Dominic Hogan (Commonwealth Scientific and Industrial Research Organisation, CSIRO), Heather Leasor (Australian Data Archive, ADA, ANU), Simon Oliver (Digital Earth Australia, DEA), Martin Schweitzer (Bureau of Meteorology, BoM), Simon O'Toole (Australian Astronomical Observatory, AAO).

Germany: Kirsten Elger and Damian Ulbricht (Helmholtz Centre Potsdam – GFZ German Research Centre for Geosciences)

USA: Robert R. Downs (Columbia University), Leslie Hsu (United States Geological Survey, USGS), Paul Jessop (International DOI Foundation), Dave Jones (StormCenter Communications Inc.), Danie Kinkaide (Biological and Chemical Oceanography Data Management Office, BCO-DMO), Benno Lee (Rensselaer Polytechnic Institute). Hampapuram K. Ramapriyan (Science Systems and Applications, Inc.).

Special thanks go to the ARDC for their support throughout this project, in particular to Gerry Ryder for her analysis of the use cases.

We also like to thank our RDA Secretariat and Technical Advisory Board (TAB) Liaisons, Stefanie Kethers and Tobias Weigel respectively, for their guidance and support.

This paper was supported by the RDA Europe 4.0 project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777388. R. Downs was supported by NASA under Contract 80GSFC18C0111 for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC).

We also thank Simon Cox, Sue Cook, and two anonymous reviewers for their constructive comments that helped improve this manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHORS CONTRIBUTIONS

Jens Klump: Contributed the collecting, compiling and analysis of use cases, and contributed to writing of this paper.

Lesley Wyborn: Contributed the collecting, compiling and analysis of use cases, and contributed to writing of this paper.

Mingfang Wu: Contributed the collecting, compiling and analysis of the use cases, contributed to writing of this paper.

Julia Martin: Contributed to documentation and analysis of use cases and reviewing of document.

Robert Downs: Contributed to collection and analysis of use cases and to writing this paper.

Ari Asmi: Contributed to the analysis of use cases and to writing this paper.

AUTHOR AFFILIATIONS

Jens Klump *orcid.org/0000-0001-5911-6022*
Mineral Resources Business Unit, Commonwealth Scientific and Industrial Research Organisation, Perth, Australia

Lesley Wyborn *orcid.org/0000-0001-5976-4943*
National Research Computational Infrastructure, Australian National University, Canberra, Australia

Mingfang Wu *orcid.org/0000-0003-1206-3431*
Australian Research Data Commons, Melbourne, Australia

Julia Martin *orcid.org/0000-0001-6939-3066*
Australian Research Data Commons, Canberra, Australia

Robert R. Downs *orcid.org/0000-0002-8595-5134*
Center for International Earth Science Information Network (CIESIN), The Earth Institute, Columbia University, United States

Ari Asmi *orcid.org/0000-0003-3933-4684*
Institute of Atmospheric and Earth System Sciences, University of Helsinki, Helsinki, Finland

REFERENCES

Albertoni, R, Browning, D, Cox, SJD, Gonzalez-Beltran, A, Perego, A, Winstanley, P, Maali, F and **Erickson, JS.** 2019. *Data Catalog Vocabulary (DCAT) – Version 2 (W3C Proposed Recommendation)*. Cambridge, MA: World Wide Web Consortium (W3C). Available at *https://www.w3.org/TR/2019/PR-vocab-dcat-2-20191119/.*

Allison, DB, Brown, AW, George, BJ and **Kaiser, KA.** 2016. Reproducibility: A tragedy of errors. *Nature News*, 530(7588): 27. DOI: *https://doi.org/10.1038/530027a*

Asch, M, Moore, T, Badia, R, Beck, M, Beckman, P, Bidot, T, Bodin, F, Cappello, F, Choudhary, A, de Supinski, B, Deelman, E, Dongarra, J, Dubey, A, Fox, G, Fu, H, Girona, S, Gropp, W, Heroux, M, Ishikawa, Y, Keahey, K, Keyes, D, Kramer, W, Lavignon, J-F, Lu, Y, Matsuoka, S, Mohr, B, Reed, D, Requena, S, Saltz, J, Schulthess, T, Stevens, R, Swany, M, Szalay, A, Tang, W, Varoquaux, G, Vilotte, J-P, Wisniewski, R, Xu, Z and **Zacharov, I.** 2018. Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *The International Journal of High Performance Computing Applications*, 32(4): 435–479. DOI: *https://doi.org/10.1177/1094342018778123*

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

Ciccarese, P, Soiland-Reyes, S, Belhajjame, K, Gray, AJ, Goble, C and **Clark, T.** 2013. PAV ontology: provenance, authoring and versioning. *Journal of Biomedical Semantics*, 4(1): 37. DOI: *https://doi.org/10.1186/2041-1480-4-37*

Cudahy, T. 2012. *Satellite ASTER Geoscience Product Notes for Australia (No. EP125895)*. Canberra, Australia: Commonwealth Scientific and Industrial Research Organisation. [Last accessed 20 January 2020]. DOI: *https://doi.org/10.4225/08/584d948f9bbd1*

DataCite Metadata Working Group. 2018. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data (No. Version 4.2)*. Hannover, Germany: DataCite e.V. DOI: *https://doi.org/10.5438/bmjt-bx77*

Dataset Exchange Working Group. 2017. Dataset Exchange Working Group. *W3C Dataset Exchange Working Group*. Available at *https://www.w3.org/2017/dxwg/wiki/Main_Page* [Last accessed 20 March 2019].

Diepenbroek, M, Grobe, H, Reinke, M, Schindler, U, Schlitzer, R, Sieger, R and **Wefer, G.** 2002. PANGAEA – an information system for environmental sciences. *Computers & Geosciences*, 28(10): 1201–1210. DOI: *https://doi.org/10.1016/S0098-3004(02)00039-0*

ESIP Data Preservation and Stewardship Committee. 2019. Data Citation Guidelines for Earth Science Data, Version 2. *Earth Science Information Partners*. DOI: *https://doi.org/10.6084/m9.figshare.8441816.v1*

Fitzpatrick, B, Pilato, CM and **Collins-Sussman, B.** 2009. *Version Control with Subversion*. Sebastopol, CA: O'Reilly Media, Inc. Available at *http://svnbook.red-bean.com/* [Last accessed 11 March 2019].

Haller, A, Janowicz, K, Cox, SJD, Lefrançois, M, Phuoc, DL, Lieberman, J, García-Castro, R, Atkinson, RA and **Stadler, C.** 2018. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation | *www.semantic-web-journal.net*. *Semantic Web Journal*, online (1878). Available at *http://www.semantic-web-journal.net/content/modular-ssn-ontology-joint-w3c-and-ogc-standard-specifying-semantics-sensors-observations* [Last accessed 11 June 2018]. DOI: *https://doi.org/10.3233/SW-180320*

Hourclé, JA. 2009. FRBR applied to scientific data. *Proceedings of the American Society for Information Science and Technology*, 45(1): 1–4. DOI: *https://doi.org/10.1002/meet.2008.14504503102*

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

Klump, J, Wyborn, LAI, Downs, RR, Asmi, A, Wu, M, Ryder, G and **Martin, J.** 2020a. Compilation of Data Versioning Use cases from the RDA Data Versioning Working Group. *Research Data Alliance*. [Last accessed 24 January 2020]. DOI: *https://doi.org/10.15497/RDA00041*

Klump, J, Wyborn, LAI, Wu, M, Downs, RR, Asmi, A, Ryder, G and **Martin, J.** 2020b. *Final Report of the Research Data Alliance Data Versioning Working Group – Principles and best practices in data versioning for all data sets big and small (Working Group Final Report)*. Kensington WA, Australia: Research Data Alliance. DOI: *https://doi.org/10.15497/RDA00042*

König-Langlo, G and **Gernandt, H.** 2008. *426 ozonesonde profiles from Georg-Forster-Station (Data)*. Bremerhaven, Germany: Alfred Wegener Institute for Polar and Marine Research. [Last accessed 9 November 2010]. DOI: *http://doi.pangaea.de/10.1594/PANGAEA.547983*

Lebo, T, Sahoo, S and **McGuinness, D.** 2013. *PROV-O: The PROV Ontology (W3C Recommendation)*. Cambridge, MA: World Wide Web Consortium (W3C). Available at *http://www.w3.org/TR/2013/REC-prov-o-20130430/*

Ledford, H and **van Noorden, R.** 2020. High-profile coronavirus retractions raise concerns about data oversight. *Nature*, 582(7811): 160–160. DOI: *https://doi.org/10.1038/d41586-020-01695-w*

Lin, D, Crabtree, J, Dillo, I, Downs, RR, Edmunds, R, Giaretta, D, De Giusti, M, L'Hours, H, Hugo, W, Jenkyns, R, Khodiyar, V, Martone, ME, Mokrane, M, Navale, V, Petters, J, Sierman, B, Sokolova, DV, Stockhause, M and **Westbrook, J.** 2020. The TRUST Principles for digital repositories. *Scientific Data*, 7(1): 144. DOI: *https://doi.org/10.1038/s41597-020-0486-7*

Mehra, MR, Ruschitzka, F and **Patel, AN.** 2020. Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet*, 395(10240): 1820. DOI: *https://doi.org/10.1016/S0140-6736(20)31324-6*

National Aeronautics and Space Administration. 2019. Data Processing Levels. *EARTHDATA*. Available at *https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels/* [Last accessed 13 July 2020].

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

Peng, RD. 2011. Reproducible Research in Computational Science. *Science*, 334(6060): 1226–1227. DOI: *https://doi.org/10.1126/science.1213847*

Preston-Werner, T. 2013. Semantic Versioning 2.0.0. *Semantic Versioning*. Available at *https://semver.org/spec/v2.0.0.html* [Last accessed 7 March 2019].

Rauber, A, Asmi, A, van Uitvanck, D and **Pröll, S.** 2016. *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) (Technical Report)*. Denver, CO: Research Data Alliance. [Last accessed 21 September 2017]. DOI: *http://doi.org/10.15497/RDA00016*

Razum, M, Schwichtenberg, F, Wagner, S and **Hoppe, M.** 2009. eSciDoc Infrastructure: A Fedora-Based e-Research Framework. In: *Research and Advanced Technology for Digital Libraries*. Heidelberg, Germany: Springer Verlag. pp. 227–238. DOI: *https://doi.org/10.1007/978-3-642-04346-8_23*

Software versioning. 2019. *Wikipedia*. Available at *https://en.wikipedia.org/w/index.php?title=Software_versioning&oldid=886437916* [Last accessed 11 March 2019].

Study Group on the Functional Requirements for Bibliographic Records. 1998. *Functional Requirements for Bibliographic Records (No. 19)*. Munich, Germany: International Federation of Library Associations and Institutions. Available at *http://www.ifla.org/publications/functional-requirements-for-bibliographic-records*. DOI: *https://doi.org/10.1515/9783110962451*

Taylor, K, Woodcock, R, Cuddy, S, Thew, P and **Lemon, D.** 2015. A Provenance Maturity Model. In: Denzer, R, Argent, RM, Schimak, G and Hřebíček, J (eds.), *Environmental Software Systems. Infrastructures, Services and Applications*. Cham, Switzerland: Springer International Publishing. pp. 1–18. [Last accessed 17 July 2015]. DOI: *https://doi.org/10.1007/978-3-319-15994-2_1*

Wilkinson, MD, Dumontier, M, Packer, AL, Gray, AJG, Mons, A, Gonzalez-Beltran, A, Waagmeester, A, Baak, A, Brookes, AJ, Evelo, CT, Mons, B, Persson, B, Goble, C, Schultes, E, van Mulligen, E, Aalbersberg, IjJ, Appleton, G, Boiten, J-W, Dillo, I, Grethe, JS, Heringa, J, Strawn, G, Velterop, J, Bouwman, J, van der Lei, J, Kok, J, Zhao, J, Wolstencroft, K, da Silva Santos, LB, Roos, M, Thompson, M, Martone, ME, Crosas, M, Swertz, MA, Axton, M, Blomberg, N, Dumon, O, Groth, P, 't Hoen, PAC, Wittenburg, P, Bourne, PE, Rocca-Serra, P, van Schaik, R, Finkers, R, Hooft, R, Kok, R, Edmunds, S, Lusher, SJ, Sansone, S-A, Slater, T, Sengstag, T, Clark, T and **Kuhn, T.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: *https://doi.org/10.1038/sdata.2016.18*

Wing, JM. 2019. The Data Life Cycle. *Harvard Data Science Review*, 1(1): 6. DOI: *https://doi.org/10.1162/99608f92.e26845b4*

TO CITE THIS ARTICLE:
Klump, J, Wyborn, L, Wu, M, Martin, J, Downs, RR and Asmi, A. 2021. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*, 20: 12, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-012*

Submitted: 17 July 2020
Accepted: 15 January 2021
Published: 23 March 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.