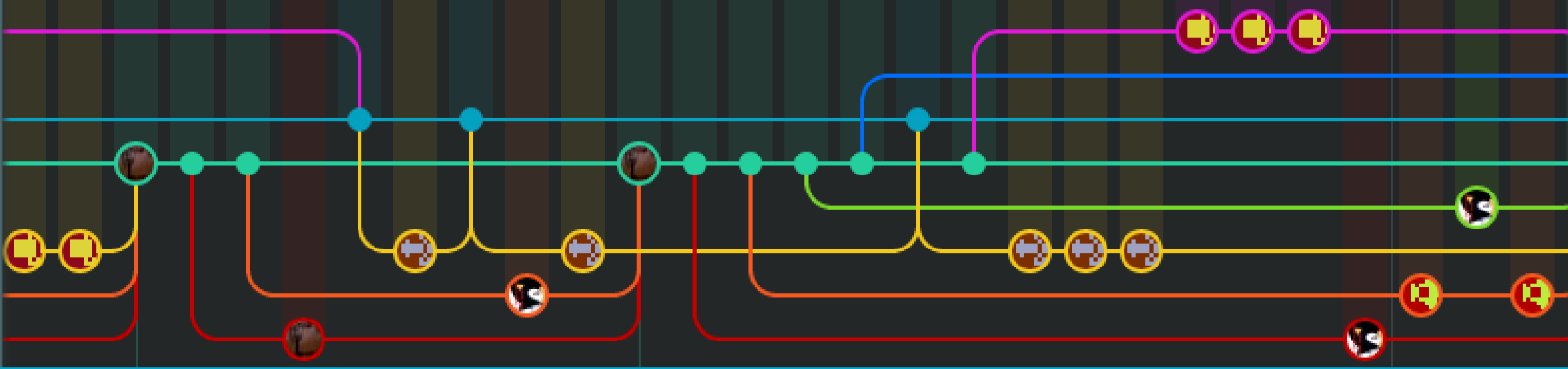


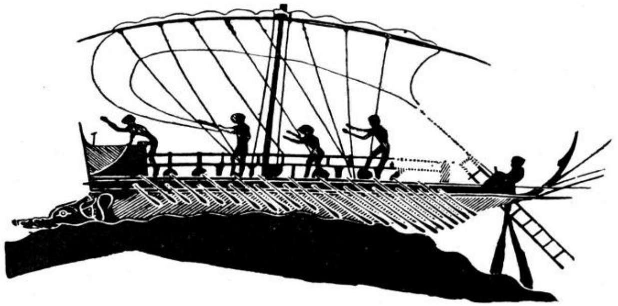
Patterns and Principles for Versioning of Research Data

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

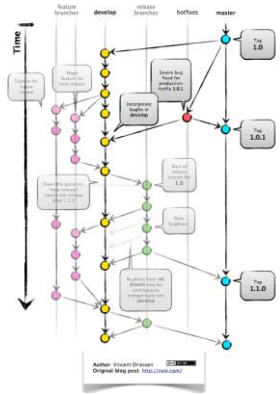
23 October 2019 | eResearch Australasia 2019 Brisbane

What is a version?

Since the beginning of the Common Era we have thought about the meaning of identity and copy (Ship of Theseus Paradox, Plutarch, 75 CE).



We also readily talk about versions, in particular in software development.


feature branches
develop
release branches
hotfixes
master
Time
Tag 1.0
Feature for future release
Major feature for next release
Severe bug fixed for production: hotfix 1.0.1
Incorporate bugfix in develop
Tag 1.0.1
Start of release branch for 1.0
From this point on, "next release" means the release after 1.1.0
Only bugfixes!
Bugfixes from rel. branch may be continuously merged back into develop
Tag 1.1.0
Author: Vincent Driessen
Original blog post: http://nvie.com/

"Version" is a fundamental concept in data and software management. But what do we mean by a "version"?

RDA Data Versioning Working Group

P8 Denver (Sept 2016): BoF on Data Versioning

P9 Barcelona (April 2017): Constituting the Data Versioning IG

P10 Montreal (Sept 2017): Data Versioning IG session

P11 Berlin (March 2018): Data Versioning WG first meeting

P12 Gaborone (Nov 2018): Data Versioning WG working meeting

P13 Philadelphia (April 2019): Data Versioning WG draft report and recommendations

P14 Helsinki (Sept 2019): Data Versioning WG final report and recommendations, TAB adoption.

Data versioning use cases

Use cases sourced from:

W3C, RDA Data Citation WG, RDA Data Foundations and Terminology IG, DA|RA, DIACHRON, BCO-DMO, NASA, USGS, NCBI/EBI

Australian Bureau of Meteorology, Geoscience Australia, Australian Integrated Marine Observation System, Australia Astronomy Observatory Data Centre, Digital Earth Australia, National Computational Infrastructure, CSIRO, National Library of Australia, EMBL-ABR, Australian Research Data Commons

Inconsistency in data versioning practices

- Can a researcher who used an older version be confident that a newer version with minor changes is not going to change their research outcome?
- Data producers use various terms for versioning: Version, Collection, Revision, Release, Edition, ... What are the differences between these terms?
- Will updating metadata count as a new version?
- How well should version history be documented to provide provenance information (e.g. document every change or only significant changes)?
- What is a significant change?
- What if only the format changes?

Inconsistency in data versioning practices

- Should a landing page of a collection/dataset (with multiple versions) point to the latest version, all published versions, all published and archived versions?
- What version related information should be recommended in data citation (version number, data-access-URL, date-of-access, etc.)?
- Should be a version number (and/or year) be added to the data title as a general practice [to improve discoverability and identification]?

Size doesn’t matter

The volume of change is not the same as the significance of the change.

Example: the edit distance of moving the decimal point in a string is small, but the consequences can be quite significant.

d = 10345.6 m vs d = 1034.56 m

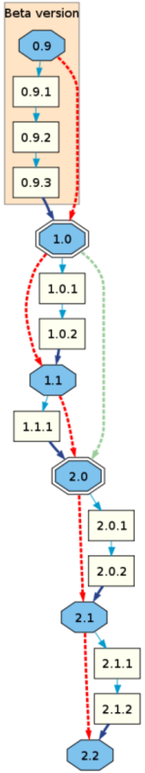
In isolation, this example seems obvious, but it might have an even less significant edit distance in the context of a large data table.

Semantic Versioning

Commonly used: Semantic Versioning

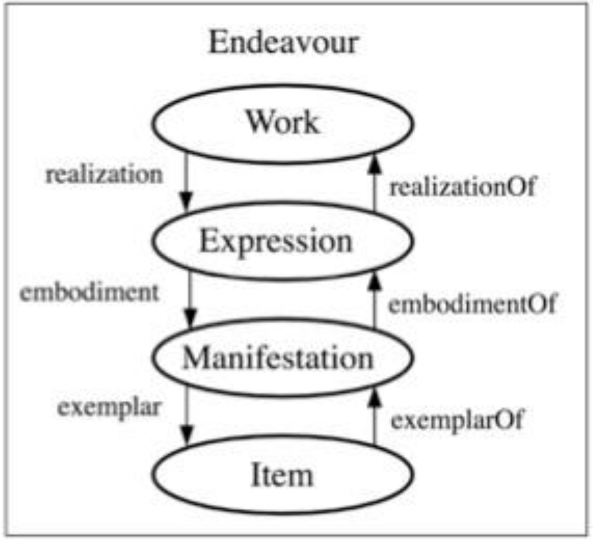
Given a version number MAJOR.MINOR.PATCH, increment the:

1. MAJOR version when you make incompatible API changes,
2. MINOR version when you add functionality in a backwards-compatible manner, and
3. PATCH version when you make backwards-compatible bug fixes.



https://semver.org/

FRBR - Functional Requirements for Bibliographic Records



Work is a ‘distinct intellectual or artistic creation’

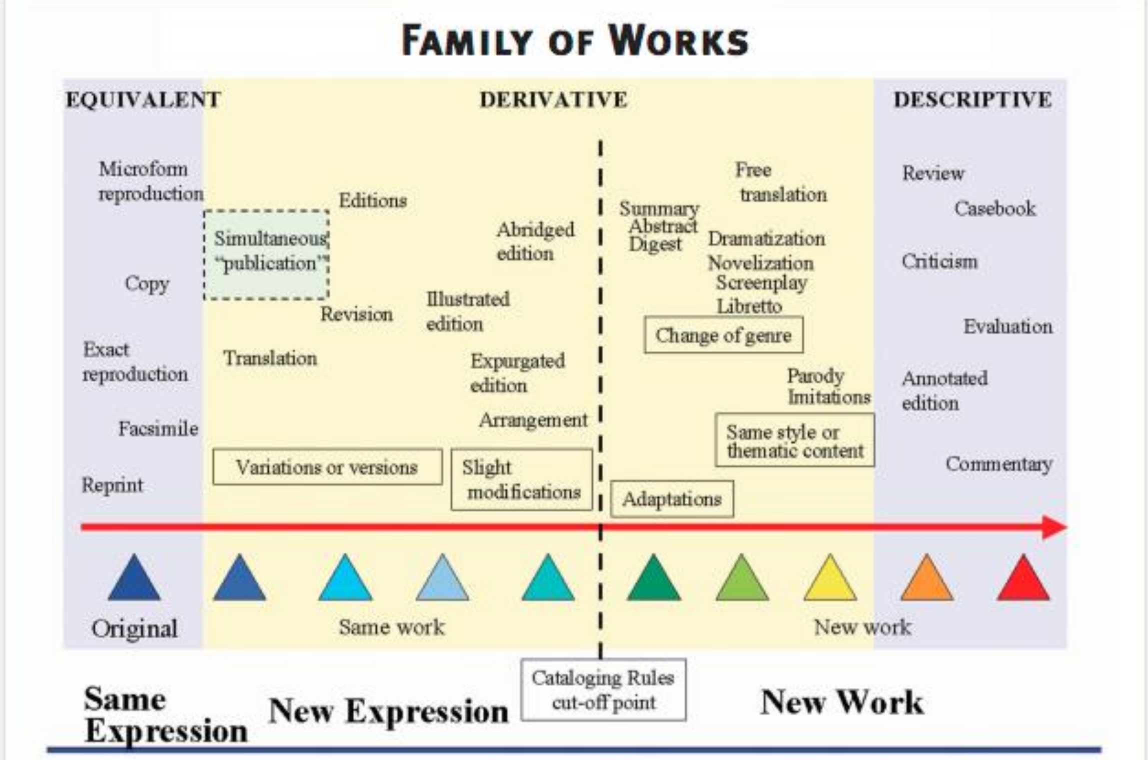
Expression is ‘the specific intellectual or artistic form that a work takes each time it is 'realized.’’

Manifestation is ‘the physical embodiment of an expression of a work’.

- Each form of dataset (e.g. CSV, figure, table, RDB) is a manifestation

Item is ‘a single exemplar of a manifestation as a concrete entity.’

From Derivative to New Work



Relationships in the Organization of Knowledge, edited by Carol A. Bean and Rebecca Green, 2001, p. 23, "Bibliographic Relationships" by Barbara B. Tillett, Figure 2, © 2001 Kluwer Academic Publishers Boston, with kind permission of Kluwer Academic Publishers.

See more in B. Tillet,2003: What is FRBR https://www.loc.gov/cds/downloads/FRBR.PDF

Versioning Patterns

- Version Control (Revision)
  - Identify each change
- Data Production (Release)
  - Communicate the significance of the change
- Objects and Collections (Granularity)
  - Identify single objects vs. collections of objects
- Formats (Manifestation)
  - Identify different formats of the same work
- Derived Products (Provenance)
  - Information about how this object was derived from a precursor.

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.

- Does a revision require the minting of a new persistent identifier?
- Should the revision be encoded in the dataset’s persistent identifier?

Identifiers for Dataset Revisions

The production of a revised dataset produces a new entity with a new identity. 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. DataCite recommends against the use of mnemonics in identifiers.

Communicate the revision to the user through the dataset’s catalogue entry.

Identifying Releases of Datasets

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**.

Identification of Data Collections

Data may be aggregated into **collections** or **timeseries**. These collections can be seen as “works of works”, similar to a journal series.

The collection (work of works) should be identified and versioned, and so should be its constituent datasets (works).

Entire time series should be identified, as should be time-stamped revisions, if the series is updated frequently. 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.

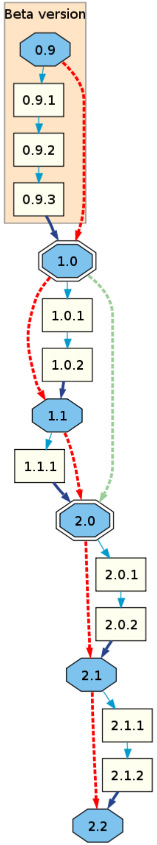
There might be technical considerations such as machine actionability that merit a machine actionable identification of different manifestations of a work, e.g. through persistent identifiers.

Provenance of a Dataset

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, etc., 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 “version” to record the “version” (release) of a dataset.

Creator (PublicationYear): Title.[Version]. Publisher. [ResourceType]. Identifier

DataCite recommends to use semantic versioning and recommends to issue a new identifier with major releases.

DataCite recommends to 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. Note that this is only the minimum required for data citation by DataCite.

Updating the metadata does not create a new version, it only changes the catalogue entry.

What’s a Version?

- Every change in the bitstream is a **revision**.
- An editorial process leads to the **release** of a data product.
- Different **manifestations** of a dataset might have different checksums but represent the same **expression** of a **work**.
- The same principles apply to collections and time series.

Key recommendations:

- Be clear about which dataset is to be identified,
- Communicate the significance of the change to designated user community of this dataset.

Thank you!

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.

Special thanks go to the ARDC for their support, in particular to Mingfang Wu, Gerry Ryder and Julia Martin.

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

ARDC Australian Research Data Commons

NCI

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

RDA RESEARCH DATA ALLIANCE

CSIRO

Thank you

Mineral Resources
Jens Klump
Team Leader Geoscience Analytics

t +61 8 6436 8828
e jens.klump@csiro.au
w people.csiro.au/Jens-Klump

Australian Research Data Commons
Mingfang Wu

e mingfang.wu@ardc.edu.au

Australian Research Data Commons
Gerry Ryder

e gerry.ryder@ardc.edu.au

Australian Research Data Commons
Julia Martin

e julia.martin@ardc.edu.au

Australian National University/NCI
Lesley Wyborn

e lesley.wyborn@anu.edu.au

Columbia University of New York
Robert Downs

e rdowns@ciesin.columbia.edu

University of Helsinki
Ari Asmi

e ari.asmi@helsinki.fi