Data Versioning

From Principles to Practical Recommendations

eResearch Australasia 2025

Jens Klump, Heinz Pampel, Mingfang Wu, Laura Rothfritz, Dorothea Strecker, Lesley Wyborn

21 October 2025

Australia's National Science Agency

I would like to begin by acknowledging the Traditional Owners of the lands that we're meeting on today, and pay my respect to their Elders past and present.

'Eternal Wisdom, Infinite Innovation'
artwork by Rachael Sarra, working with Gilimbaa.

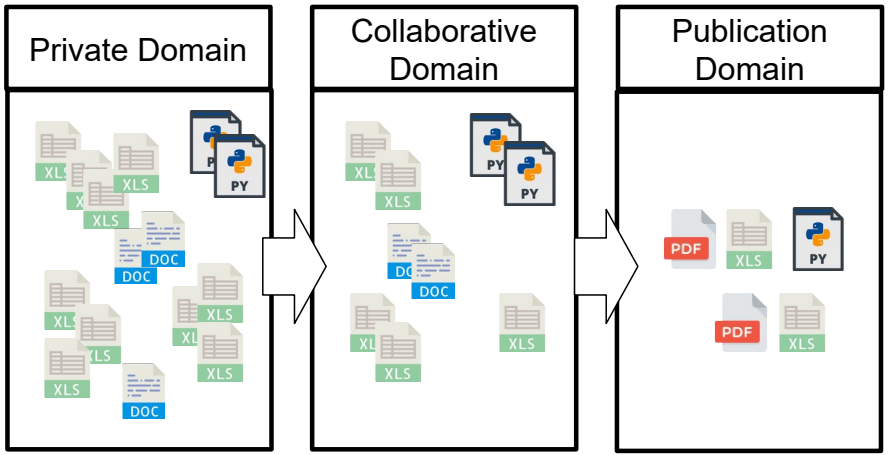
What is a Version?

When we compare two digital objects, it is easy to determine whether they are identical by looking at the bitstream.

- But what does this difference mean?
- Same format, different content?
- Same content, different format?
- What is important about this difference?

- Implications for data management?
- What, who and how do I cite?
- When do I need a new DOI?

The Life Cycle is Not Linear



Many digital artefacts are created in the process from conception of a research idea to the publication of data and their interpretation in one or more research papers.

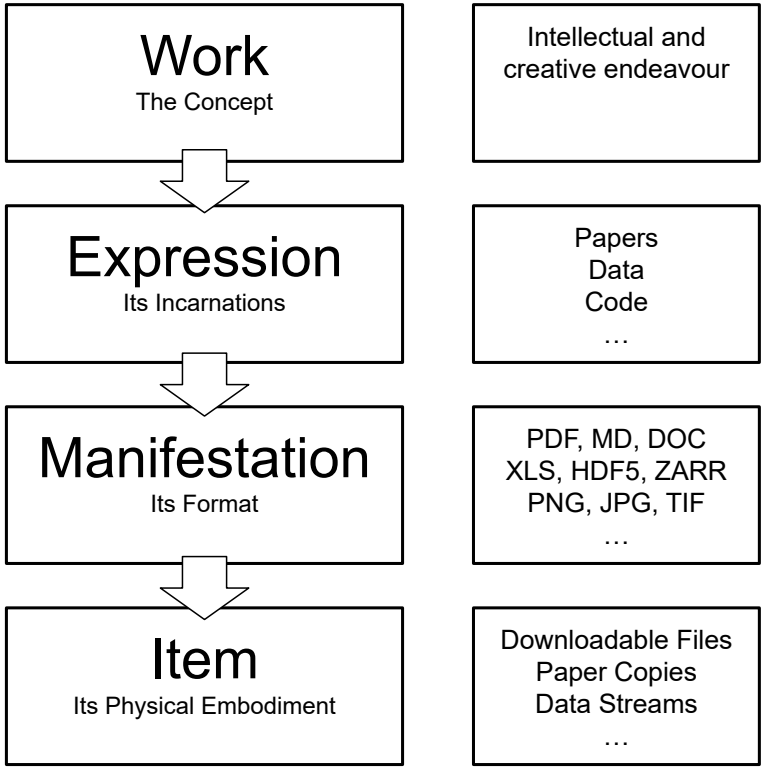
What we see is not the full story

The paper is the public-facing display (“storefront”) of the research, showcasing the interpretations and insights.

The data is the underlying raw material that justifies those interpretations.

The paper acts as the bridge between the data and the broader understanding it supports.

From Idea to Download



To be able to describe all the patterns how objects in the research life cycle relate to each other, we developed a new terminology.

We adopted the FRBR terminology as a framework to develop actionable recommendations.

Revision vs. Release

Revision

- Changes to the bitstream
- Can be automated

- No indication of significance of changes

Release

- Editorial/curatorial process
- Cannot be automated

- Communicating the significance of the changes to the users is essential.

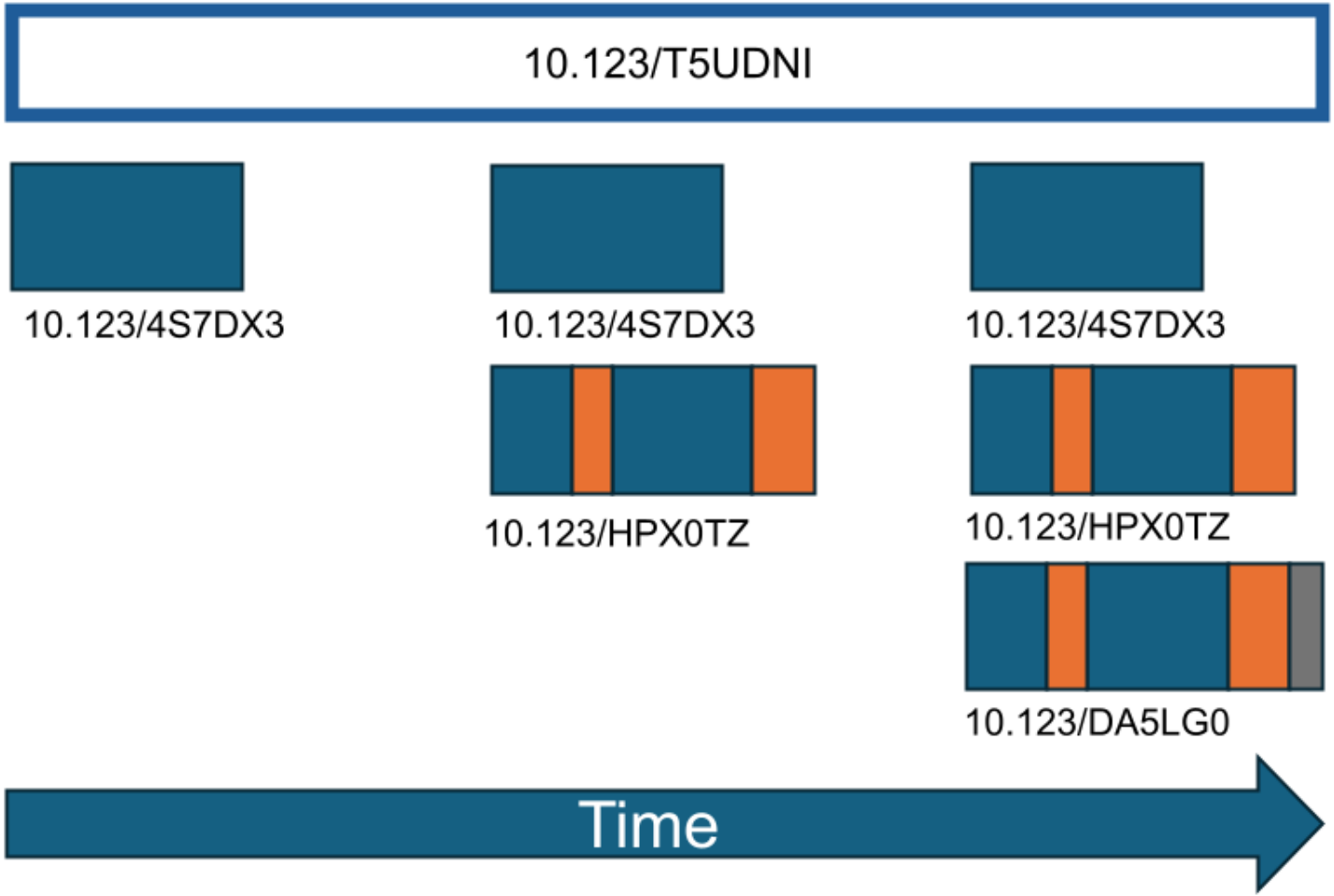
What is a significant change?

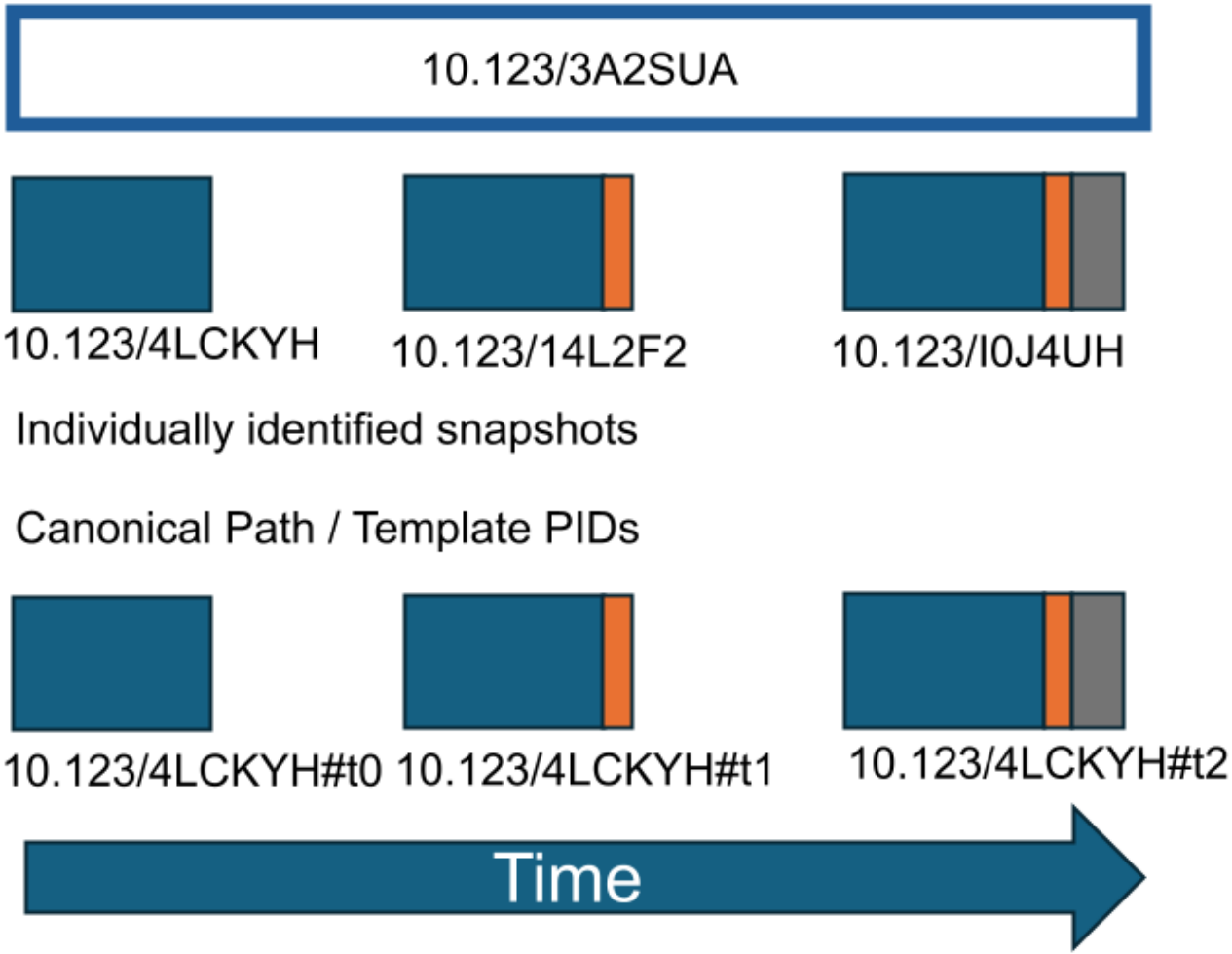
3.14 or 31.4?

- Small edit distance, but potentially a large change.

ὁμοούσιος or ὁμοιούσιος?

- Small change
- Resulted in the first major schism of the Christian church in 325 CE.
- Not very relevant anymore, except for some fringe denominations.


10.123/T5UDNI
10.123/4S7DX3
10.123/4S7DX3
10.123/4S7DX3
10.123/HPX0TZ
10.123/HPX0TZ
10.123/DA5LG0
Time


10.123/3A2SUA
10.123/4LCKYH
10.123/14L2F2
10.123/l0J4UH
Individually identified snapshots
Canonical Path / Template PIDs
10.123/4LCKYH#t0
10.123/4LCKYH#t1
10.123/4LCKYH#t2
Time

10.123/OW058J

10.123/
RRU5LE

10.123/
L81HFN

10.123/
JBV5PR

Mirroring and Re-publication

- Mirroring and re-publication refer to making data available on another platform.
- The mirrored or re-published data should always refer to the original source (authoritative version) using the “IsIdenticalTo” attribute in the DataCite metadata. “isAuthoritative”?
- This explicit relationship helps to manage multiple items that embody the same Manifestation, e.g. same GeoTIFF on multiple portals.

Loss, Deletions and Retractions

- **Data unavailable:** A persistent identifier should always resolve to a landing page displaying the metadata, even when a dataset has become unavailable.
- **Metadata unavailable:** If both the data and the metadata have become unavailable, their persistent identifier should resolve to a tombstone page.
- **Retractions:** Display the metadata with an appropriate release note describing that the data have been retracted.

Versioning of Metadata

- The rise of machine-readable metadata records ingested by metadata aggregators creates a new use case for metadata identification and versioning.
- Human users and automated systems must be able to distinguish between different versions of a metadata record and identify the authoritative source.
- Communicate clearly to users which identifier refers to the object described by the metadata and which identifier refers to the metadata record itself.

How to Cite Data

- Citation of data follows the established practices for citation in scholarly communications, even though not all journals might yet accept it.
- Credit is given to the creator of a data product by citing the **Expression**.
- Citing the Item will credit the infrastructure providing access to the data product, but in the case of mirrored data, it might lead to skewed metrics.

Summary

- We created a framework and terminology to describe how data relate to other elements of scholarly communication.
- We used this framework to develop actionable recommendations for data versioning.
- The recommendations were discussed with data practitioners in a series of workshops.
- The recommendations were published as a report for the Berlin University Alliance, and a peer-reviewed journal publication in preparation. doi:10.5281/zenodo.13743876

Thank you!

Jens Klump (CSIRO)
jens.klump@csiro.au

Heinz Pampel (HU Berlin)
heinz.pampel@hu-berlin.de

Mingfang Wu (ARDC)
mingfang.wu@ardc.edu.au

Laura Rothfritz (HU Berlin)
laura.rothfritz@hu-berlin.de

Dorothea Strecker (HU Berlin)
dorothea.strecker@hu-berlin.de

Lesley Wyborn (ANU NCI, ARDC)
lesley.wyborn@anu.edu.au

Berlin University Alliance

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Australian Research Data Commons

doi:10.5281/zenodo.13743876