


Data Versioning WG

RDA Plenary 11 | 23 March 2018

Agenda

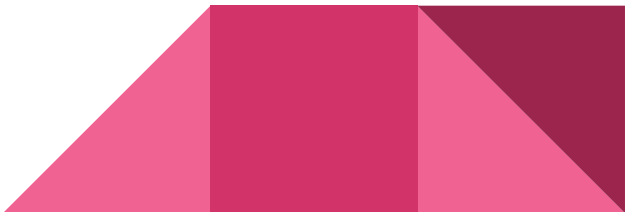
- Introduction
 - Recap of Why, How and What of Data Versioning
 - Review of use cases, including the W3C Dataset Exchange Use Cases and Requirements
 - Work plan for RDA Data Versioning WG
 - Engagement with other RDA and external groups
 - Outline of white paper on data versioning practices
 - Scheduling of online meetings up to Plenary 12
- 

Meeting Objectives

- Establish a work plan for this RDA Working Group on developing agreed practices for Data Versioning. This includes planning of how to engage with other groups in RDA and externally where data versioning is required.
- Further documented cases where groups and organisations are undertaking data versioning.
- Develop the outline of a white paper on recommendations for versioning for a spectrum of data types (files, databases, unstructured data, model runs, etc.), and align these with the practices for the assignment of persistent identifiers.

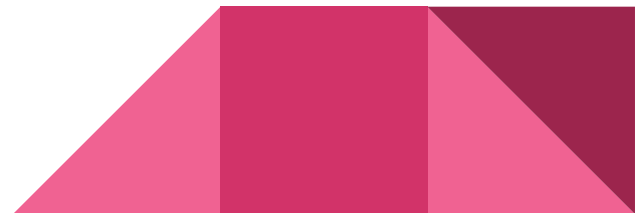


RDA Guidelines: Review Criteria for a Working Group

- **Focus:** Are there measurable outcomes?
 - **Impact and Engagement:** Will the outcome(s) of the Working Group be taken up by the intended community? Will the outcome(s) of the Working Group foster data sharing and/or exchange?
 - **Timeframe:** Can the proposed work, outcomes /deliverables, and Action Plan described in the Case Statement be accomplished in 12-18 months?
 - **Scope/Fit:** Is the scope too large for effective progress, too small for an RDA effort, or not appropriate for the RDA?
- 

Recap: The Why, How and What

- Datasets published on the Web may change over time.
- Some datasets are updated on a scheduled basis, and other datasets are changed as improvements in collecting the data make updates worthwhile.
- Others are updated because errors are found
- In order to deal with these changes, new versions of a dataset may be created.
 - What is the significance of the change?
 - Is the new version compatible with the previous version?



FRBR and Provenance

What is the authoritative copy?

Version becomes a question of provenance, instance, identity

This becomes relevant when data sets are transferred to other repositories or derivative data sets are produced.





NCI
AUSTRALIA

Issues with 'external' data versioning you may not of thought of ... yet (with apologies to Beethoven)

Lesley Wyborn

National Computational Infrastructure ANU

We need to move on from the 'book on the shelf' mentality

The traditional thought mode for finding a dataset via metadata is the old library way of locating the 'book on the shelf' via the card catalogue. File is catalogued, found and down loaded for local processing: curation is minimal is this mode appropriate for access for datasets via services?



http://commons.wikimedia.org/wiki/File:Shelves_of_Language_Books_in_Library.JPG



http://en.wikipedia.org/wiki/Library_catalog#/media/File:Schlagwortkatalog.jpg

1. Web Services

2. Analysis Ready Data (ARD)

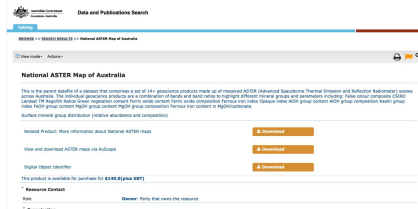
- Products that make data more accessible, easier to analyze, and reduce the amount of time users spend on data processing prior to analysis
- i.e., Data are consistently processed to the highest scientific standards and level of processing required for common access
- May not be all that big

3. High Performance Data (HPD)

- HPC variation of ARD
- Data moved to be close to compute because bandwidth limits capacity to access it in realistic time frames
- Multiple individual data sets, potentially from multiple sources, aggregated into homogenous data sets to enable 'high performance' access, including parallelisation to improve access



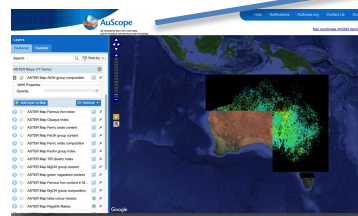
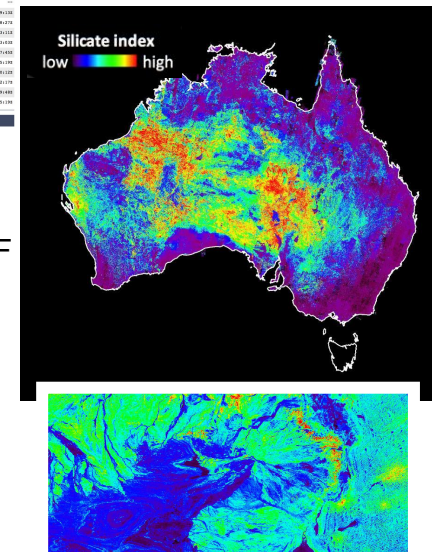
CSIRO ASTER Collection
Data in TIFF format, and
the files are broken up into
chunks of ~2GB for
downloading: there are
1258 files.



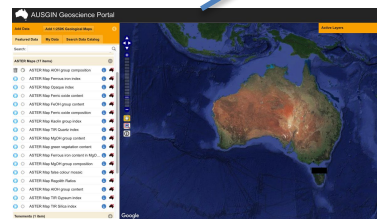
GA ASTER Collection
on external hard drives
in either BSW or
GeoTIFF formats:
posted to clients at a
cost of \$154.00



NCI ASTER Collection
available as 25 files in netCDF
(10 files are 60 GB)
set up for HPD in-situ access
by OGC web services as
national seamless coverages



AuScope Portal (WMS)



AusGIN Portal (WMS)

- This is a **conceptual** model developed by the International Federation of Library Associations and Institutions (IFLA)
- It represents a more holistic approach to retrieval and access of book resources
- The ways that people can use FRBR data have been defined as follows:
 - to find entities in a search
 - to identify an entity as being the correct one
 - to select an entity that suits the user's needs
 - or to obtain an entity (physical access or licensing)

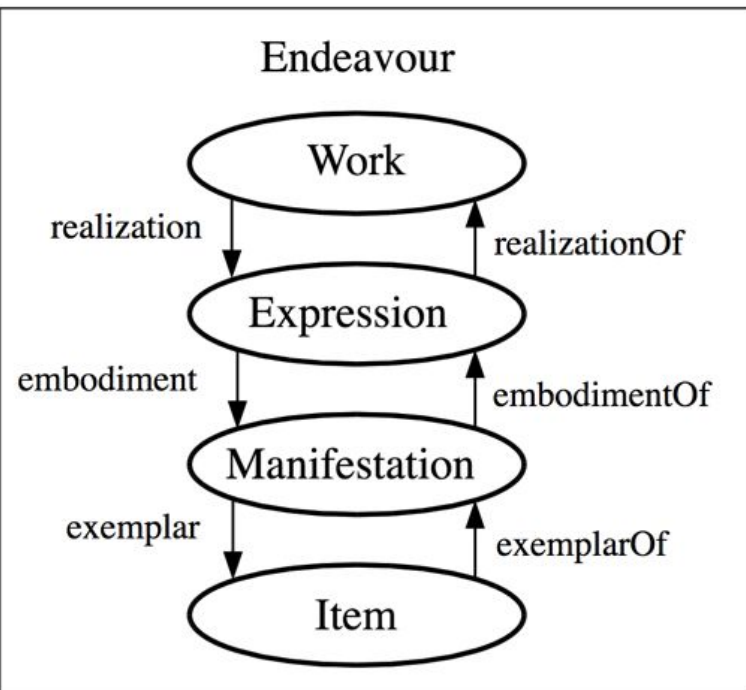
Source: https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records

B. Tillet, 2003: What is FRBR <https://www.loc.gov/cds/downloads/FRBR.PDF>

- FRBR comprises 3 groups of entities:
 - **Group 1** entities are work, expression, manifestation, and item (WEMI). They represent the products of intellectual or artistic endeavor.
 - **Group 2** entities are person, family and corporate body, responsible for the custodianship of Group 1's intellectual or artistic endeavor.
 - **Group 3** entities are subjects of Group 1 or Group 2's intellectual endeavor, and include concepts, objects, events, places.

Source: https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records

B. Tillett, 2003: What is FRBR <https://www.loc.gov/cds/downloads/FRBR.PDF>



- **Group 1** entities are work, expression, manifestation & item.
 - They represent the products of intellectual or artistic endeavor.
- **A Work** is a 'distinct intellectual or artistic creation'
 - e.g., *Beethoven's Ninth Symphony*
- **Expression** is 'the specific intellectual or artistic form that a work takes each time it is 'realized.'
 - e.g., *Each draft score of the Ninth that Beethoven writes*
- **Manifestation** is 'the physical embodiment of an expression of a work'.
 - e.g., *The performance the London Philharmonic made of the Ninth in 1996*
- **Item** is 'a single exemplar of a manifestation. The entity defined as item is a concrete entity.'
 - e.g., *Each copy of the 1996 pressings of that 1996 recording is an item.*

Source: https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records
 B. Tillet, 2003: What is FRBR <https://www.loc.gov/cds/downloads/FRBR.PDF>

Equivalence relationships

- Exist between exact copies of the same manifestation of a work or between an original item and reproductions of it, so long as the intellectual content and authorship are preserved
 - e.g., reproductions such as copies, issues, facsimiles and reprints, photocopies, microfilms.

Derivative relationships

- Exist between a bibliographic work and a modification based on the work, e.g.:
 - Editions, versions, translations, summaries, abstracts, and digests
 - Adaptations that become new works but are based on old works
 - Genre changes
 - New works based on the style or thematic content of the work

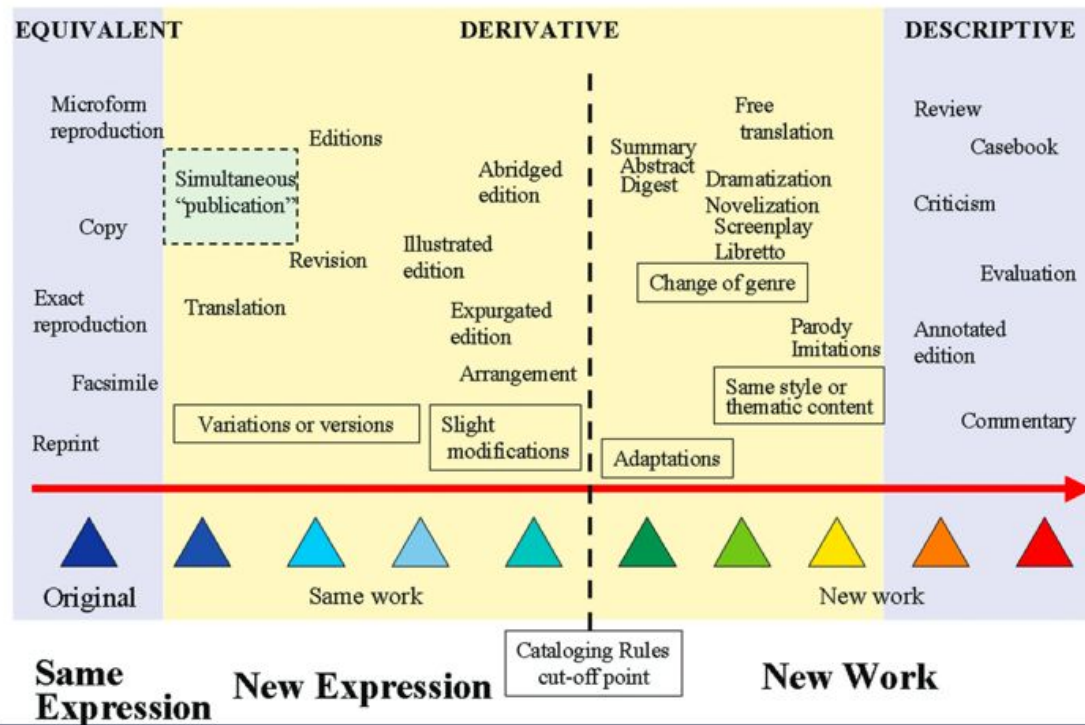
Descriptive relationships

- Exist between a bibliographic entity and a description, criticism, evaluation, or review of that entity
 - e.g., between a work and a book review describing it.
 - e.g., annotated editions, casebooks, commentaries, and critiques of an existing work.

Source: https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records

B. Tillet, 2003: What is FRBR <https://www.loc.gov/cds/downloads/FRBR.PDF>

FAMILY OF WORKS



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. Tillett, 2003: What is FRBR
<https://www.loc.gov/cds/downloads/FRBR.PDF>

Data Product Level	Description
Level 0	Reconstructed, unprocessed instrument data at original resolution, time ordered, all communications artifacts removed.
Level 1A	Level 0 data time referenced and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (i.e., platform ephemeris) computed and appended, but not applied to Level 0 data.
Level 1B	Radiometrically corrected and geolocated Level 1A data that have been processed to sensor units.
Level 1C	Level 1B data that have been spatially resampled.
Level 2	Derived geophysical parameters at the same resolution and location as the Level 1 data from which they are derived.
Level 3	Geophysical parameters derived from Level 1 or 2 data that have been spatially and/or temporally re-sampled to a global grid.
Level 4	Geophysical parameters derived by assimilating Level 1, 2, or 3 data into a land surface model.

Source: <http://smap.jpl.nasa.gov/data/>

WHAT Collection



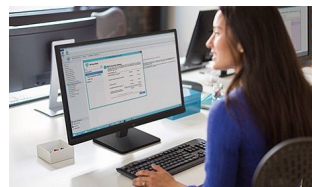
Dataset (Data Set?)



Granule



WHERE



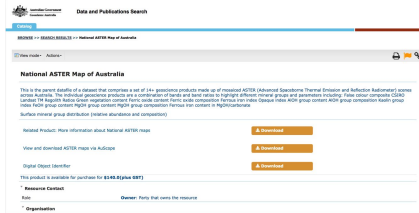
WHO



HOW

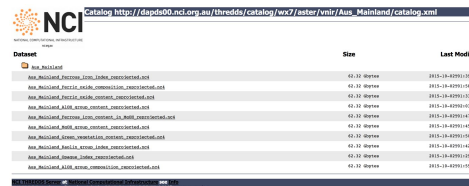


What is Beethoven's 9th – i.e., the body of work = ?JAXA, ?CSIRO, ?Geological Surveys

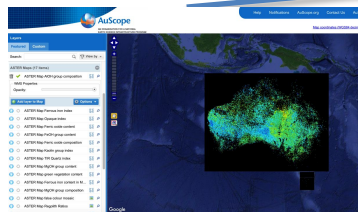
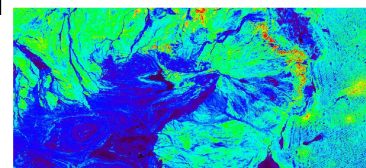
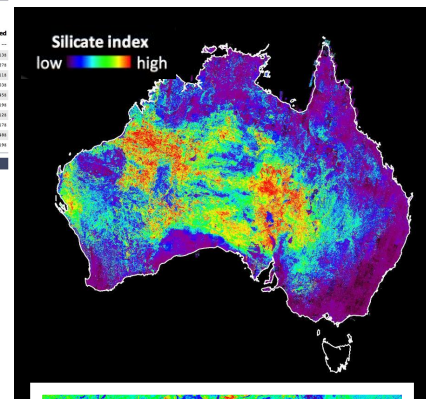


CSIRO ASTER Collection
Data in TIFF format, and
the files are broken up into
chunks of ~2GB for
downloading: there are
1258 files.

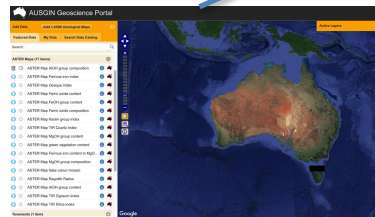
GA ASTER Collection
on external hard drives
in either BSW or
GeoTIFF formats:
posted to clients at a
cost of \$154.00



NCI ASTER Collection
available as 25 files in netCDF
(10 files are 60 GB)
set up for HPD in-situ access
by OGC web services as
national seamless coverages



AuScope Portal (WMS)



AusGIN Portal (WMS)

And then there is the issue of the push or pull copies....

Use Cases

So far the Versioing WG has compiled a list of use cases at

<https://docs.google.com/document/d/1TfBPIfjTVg0YcFuw0UzAXPYrRmyZ6PCctxKx8-uGg>

RDA Use Cases Group: <https://rd-alliance.org/groups/use-cases-group.html>

Examples came from W3C, RDA Data Citation WG, RDA Data Foundations and Terminology IG, DAIRA, DIACHRON, USGS, ANDS.



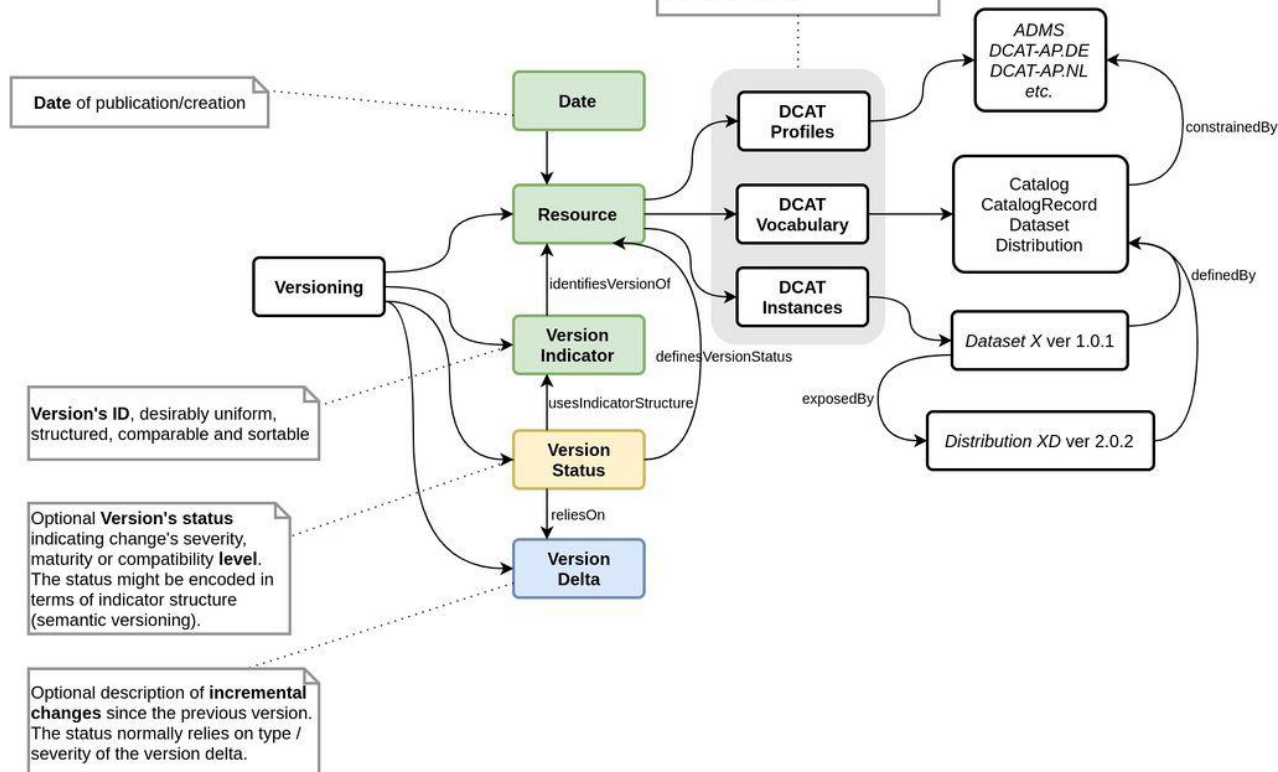
Status description

GREEN - definition mandatory

YELLOW - definition suggested

BLUE - definition optional

Type of resources that are subject to a versioning mechanism (i.e. are governed by a distinguished life-and update cycle).



Related RDA Groups

RDA Use Cases Group: <https://www.rd-alliance.org/groups/use-cases-group.html>

Pointers to the use cases of the RDA Working and Interest Groups.



Work Plan for WG Data Versioning

- March 2018 - P11 Berlin:
- November 2018 - P 12 Gaborone:
- March 2019 - P13:
- August 2019: WG ends
- September - P14: Final report





Thank you!

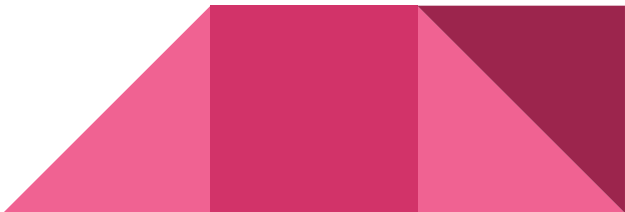
See you at RDA P12

RDA Data Versioning WG

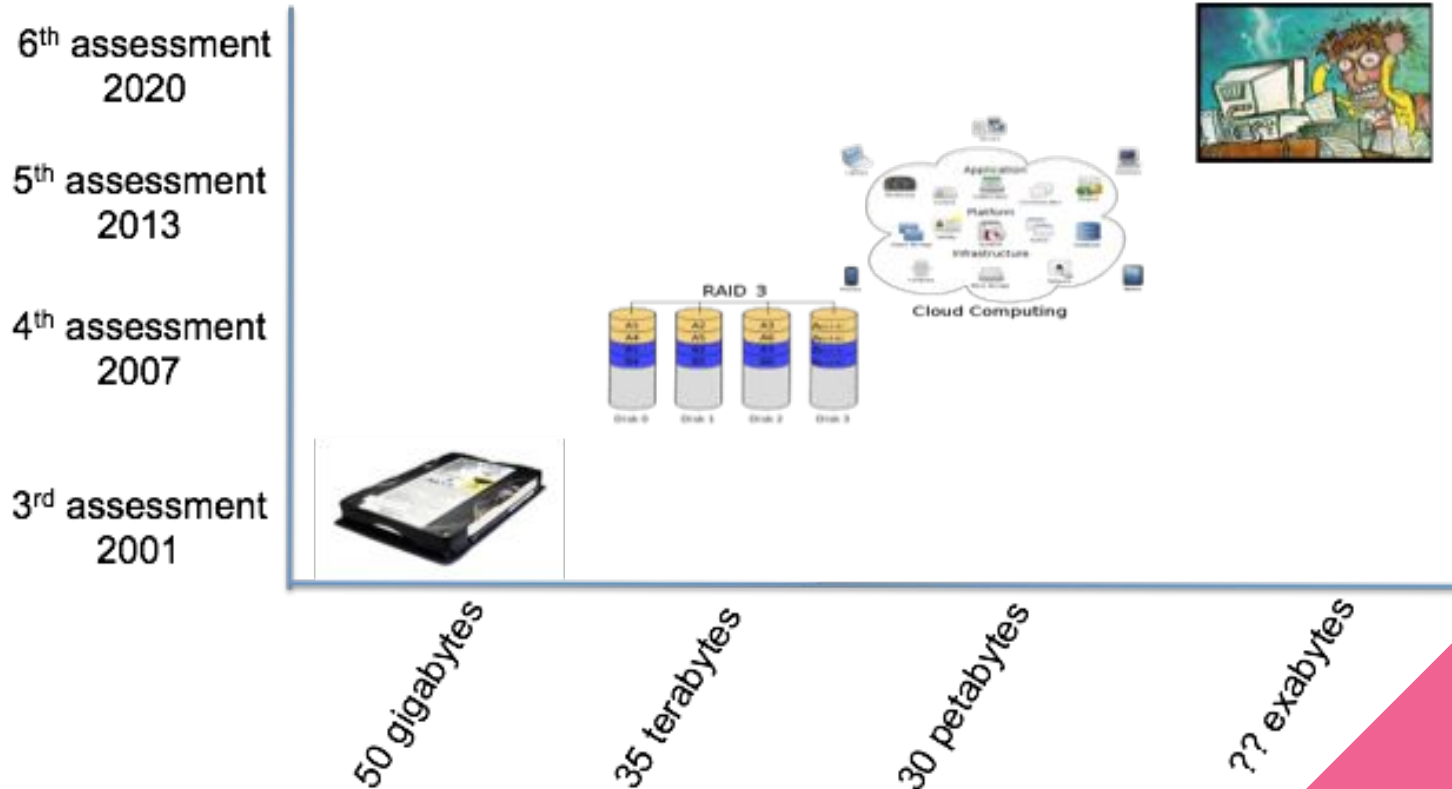
Jens Klump, Lesley Wyborn, Ari
Asmi, Robert Downs

[https://www.rd-alliance.org/groups
/data-versioning-wg](https://www.rd-alliance.org/groups/data-versioning-wg)

Agenda

- Introduction (5 min)
 - Recap of Why, How and What of Data Versioning (10 min)
 - Review of use cases, including the W3C Dataset Exchange Use Cases and Requirements (20 min)
 - Work plan for RDA Data Versioning WG (15 min)
 - Engagement with other RDA and external groups (10 min)
 - Outline of white paper on data versioning practices (20 min)
 - Scheduling of online meetings up to Plenary 12 (10 min)
- 

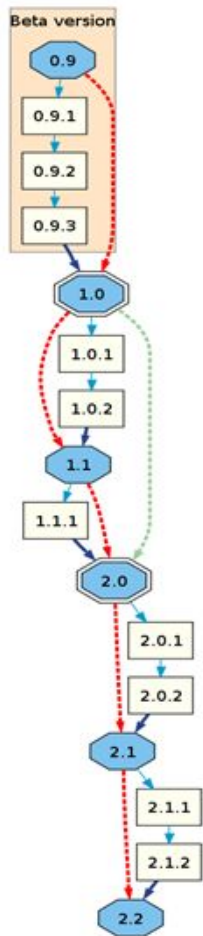
The issue will not go away



Versions vs Builds

- In version control systems, version numbers are incremented sequentially.
- Many software publication systems distinguish between “builds” (sequentially numbered) and “versions” (releases with semantic labelling).
- The number of changes is not a useful indicator of change as even small changes can have significant consequences.





Semantic Versioning

Versioning can carry meaning (Semantic Versioning)

- Sequence-based identifiers
 - Change significance
 - Degree of compatibility
 - Designating development stage

Version n.n.n : (major release).(minor release).(patch)

RDA Case Statement

<https://docs.google.com/document/d/1jZoON7biETH46lvoXcyxt0is4qyt0CWGJrVpHTyfpk>




Versioning and Identifiers

- It is currently being debated in FORCE11 and DataCite about persistently identifying versions:
 - Should versions be reflected in the formatting of DOIs?
 - How should other versions be referred to in DOI metadata?
- This is the next step once we have a common understanding of data versioning.



RDA Guidelines: Case Statement (general)

Working Groups are expected to:

- Develop clear outcomes and put them into action to create tangible progress (see also [Working Group Goals and Outcomes](#) and the [RDA Outputs and Intellectual Property Policy](#)).
 - Work openly and transparently with respect to the community.
 - Document their efforts as they operate.
 - Meet regularly with the RDA to facilitate coordination and communication.
- 

RDA Guidelines: Case Statement (general)

- What is the research case (will the WG produce something useful)?
- What is the business case (will people use it)?
- Is there capacity (are the right people involved to adopt and implement)?

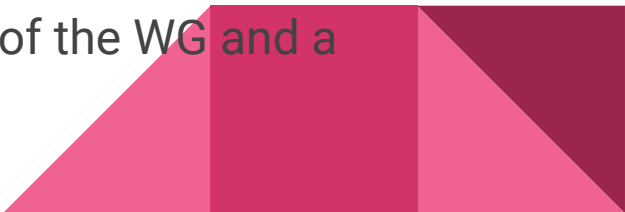


RDA Guidelines: Case Statement (format)

- WG Charter: A concise articulation of what issues the WG will address within a 12-18 month time frame and what its “deliverables” or outcomes will be.
- Value Proposition: A specific description of who will benefit from the adoption or implementation of the WG outcomes and what tangible impacts should result.
- Engagement with existing work in the area: A brief review of related work and plan for engagement with any other activities in the area.



RDA Guidelines: Case Statement (format)

- Work Plan: A specific and detailed description of how the WG will operate including:
 - Adoption Plan: A specific plan for adoption or implementation of the WG outcomes within the organizations and institutions represented by WG members, as well as plans for adoption more broadly within the community. Such adoption or implementation should start within the 12-18 month timeframe before the WG is complete.
 - Initial Membership: A specific list of initial members of the WG and a description of initial leadership of the WG.
- 

Work Plan

Objective of this session is to draft a workplan and a case statement.

