

Research data citation and versioning practices

Megan Force
Digital Research Analyst, Thomson Reuters
megan.force@thomsonreuters.com
July 12th, 2016

The intelligence, technology and human expertise
you need to find trusted answers.

Recommended practices

- Researcher/data author
- Repository/publisher/data provider
- Literature publisher/funding organization

Versioning practices

- Questions
- Study results
- Takeaways

The Data Citation Index

- **Data repositories evaluated for inclusion based on factors such as persistence, evidence of reuse, editorial content**
- **Collaboration on metadata handling with data repository partners**
- **Links from data to research literature**

Force, M et al. J Comput Aided Mol Des (2014) 28: 1043.
doi:10.1007/s10822-014-9768-5

Launched October 2012

~ 6 million data records

Repository/Source: Comprises data studies and/or data sets. Stores and provides access to the raw data.

Data Study: Descriptions of studies or experiments with associated data which have been used in the data study. Includes serial or longitudinal studies over time.

Data Set: A single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment.



Recommended practices to promote data citation and tracking

White paper available at:

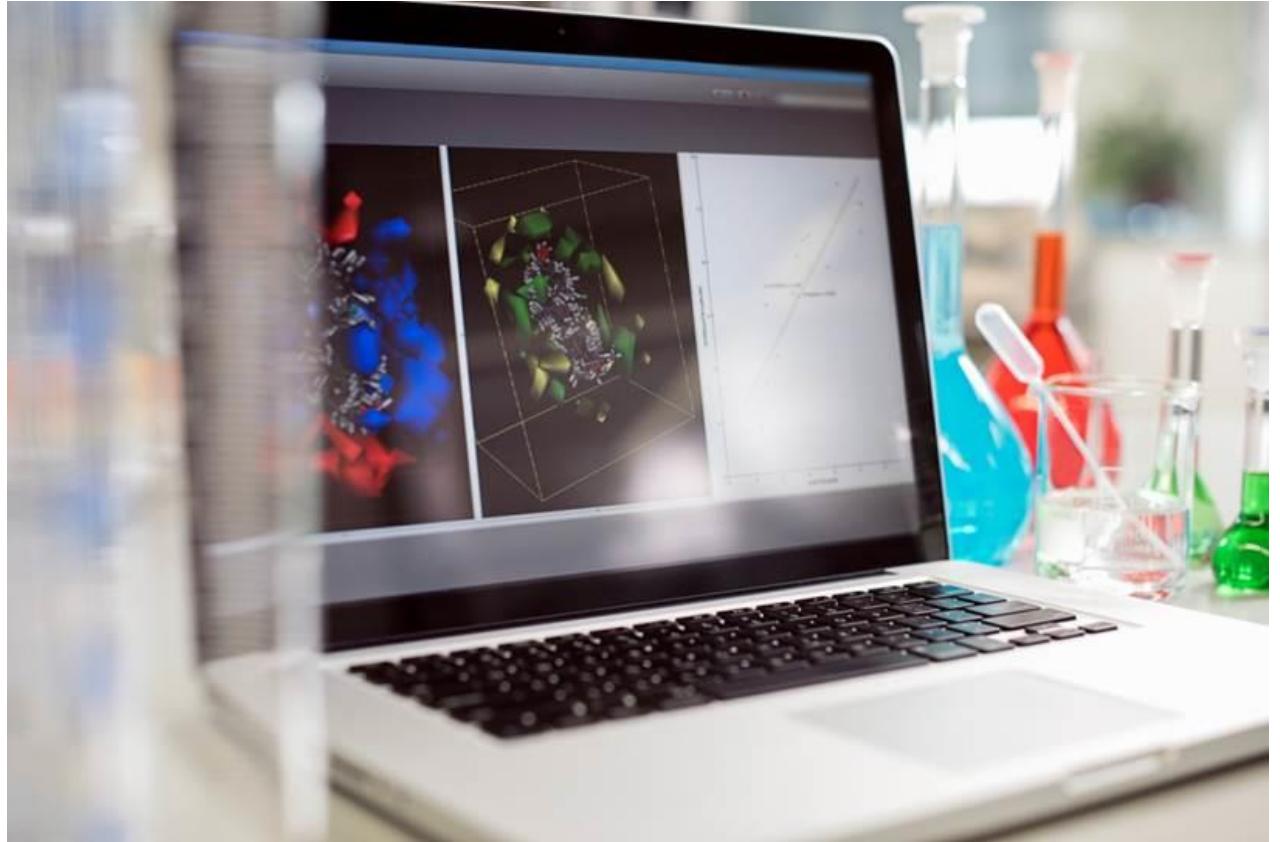
http://wokinfo.com/publisher_relations/data/

The intelligence, technology and human expertise
you need to find trusted answers.

Researcher/data author

Recommended practices

- Treat data equally with other citable research output
- Deposit data in an established data repository committed to long term preservation
- Practice detailed, formal data citation in data and publications



Researcher/data author

Recommended practices

- Provide required metadata to enable citation and discovery
 - Allows creation of a data citation using DataCite guidelines
 - Compliance with DataCite metadata schema
 - Allows matching of data citations encountered to known data records

- ❖ Unique ID in repository
- ❖ Date provided
- ❖ Author
- ❖ Repository
- ❖ URL/DOI
- ❖ Title
- ❖ Year Published
- ❖ Version

Repository/data publisher/data provider

Recommended practices

- Curate and validate metadata for completeness, accuracy, and consistency
- Issue permanent IDs for data objects
- Provide unique landing pages for data objects



Repository/data publisher/data provider

Recommended practices

- Maintain detailed update information
- Indicate data resource type in metadata
- Ensure clear attribution for data objects
- Provide metadata harvesting facility
- Establish, document, implement, and maintain policies on dataset versioning
- Document repository mission and policies for inclusion

Literature publisher/funding organization

Recommended practices

- Establish clear data management plan policies and guidelines for data deposition
- Develop formal data citation policies and clear guidance for authors on data citation formats
- Enforce requirements for data deposition and citation
- Establish metadata criteria for persistent and unique identification of cited data

Versioning practices in the Data Citation Index

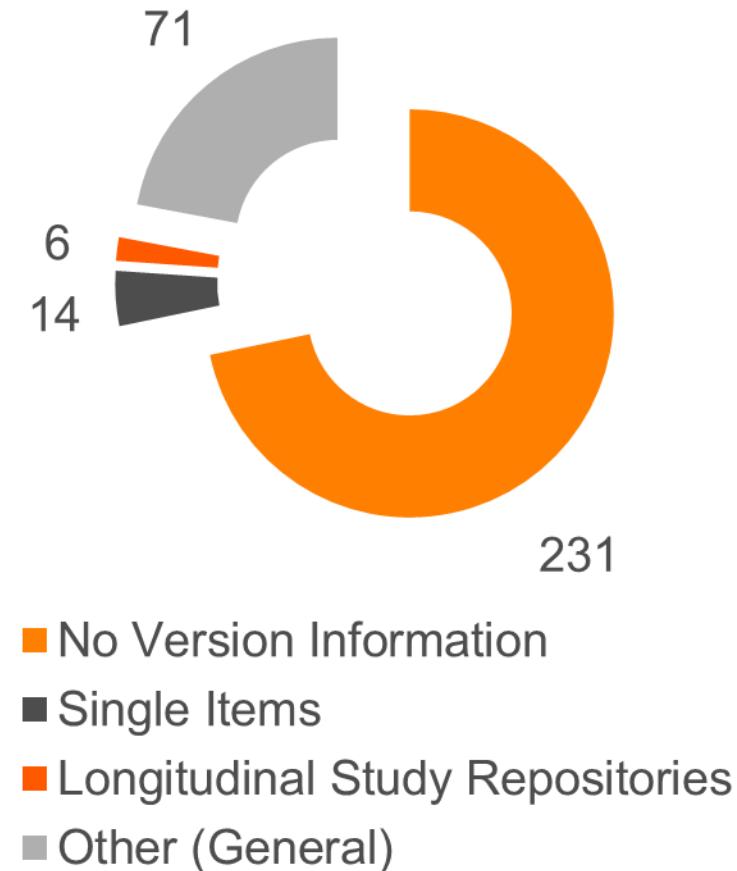
DCI Versioning Study

- 72% of all data repositories in DCI were found to have no version information for datasets
- From the 28% with version information, single item and longitudinal study subgroups with similar versioning practices were identified

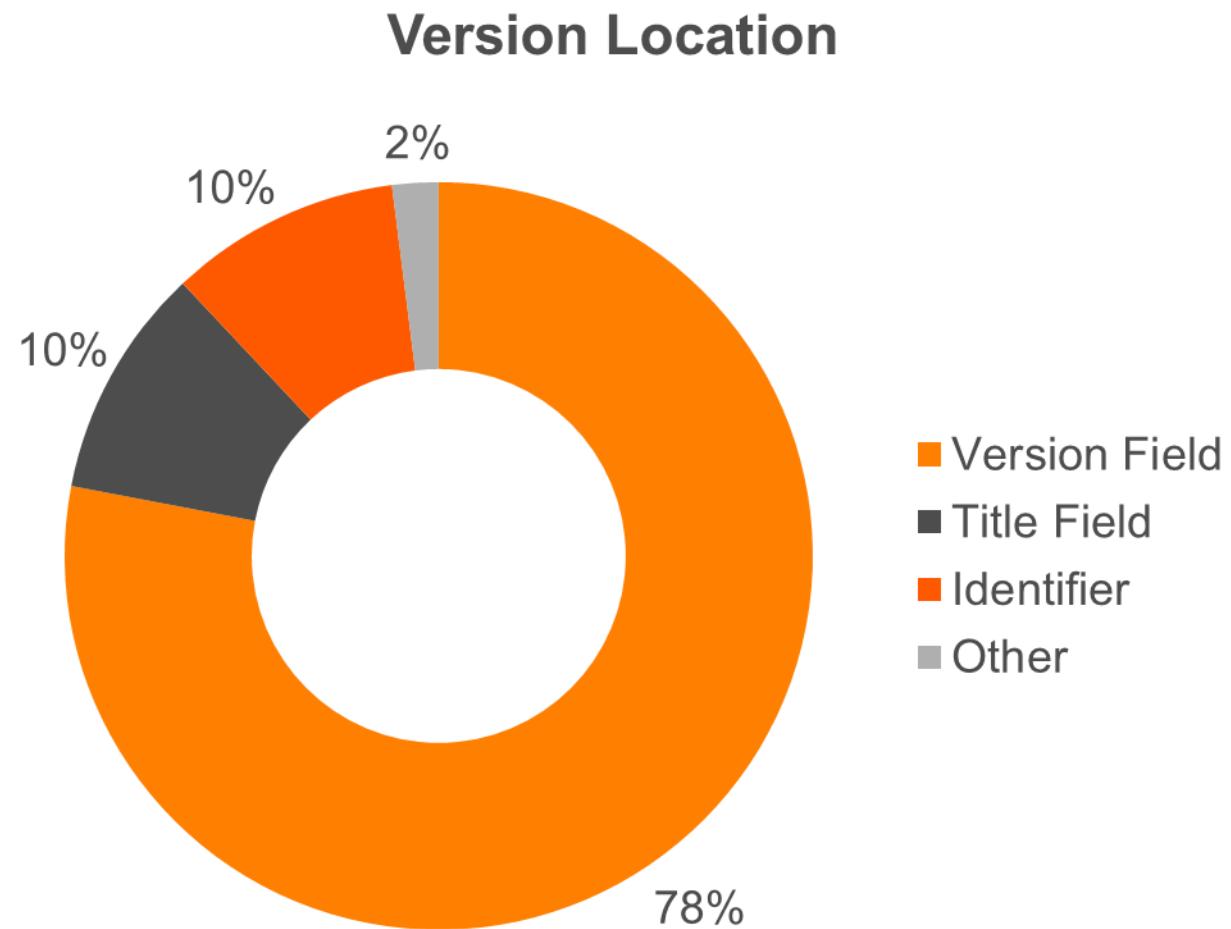
Questions:

- Where is version information found in the metadata?
- DOIs/URLs for each version?
- Are versions cited?
- Does the recommended citation include version?
- Landing page for each version?
- Does the repository retain older versions of the dataset?

June 2016: 322 Data Repositories in DCI



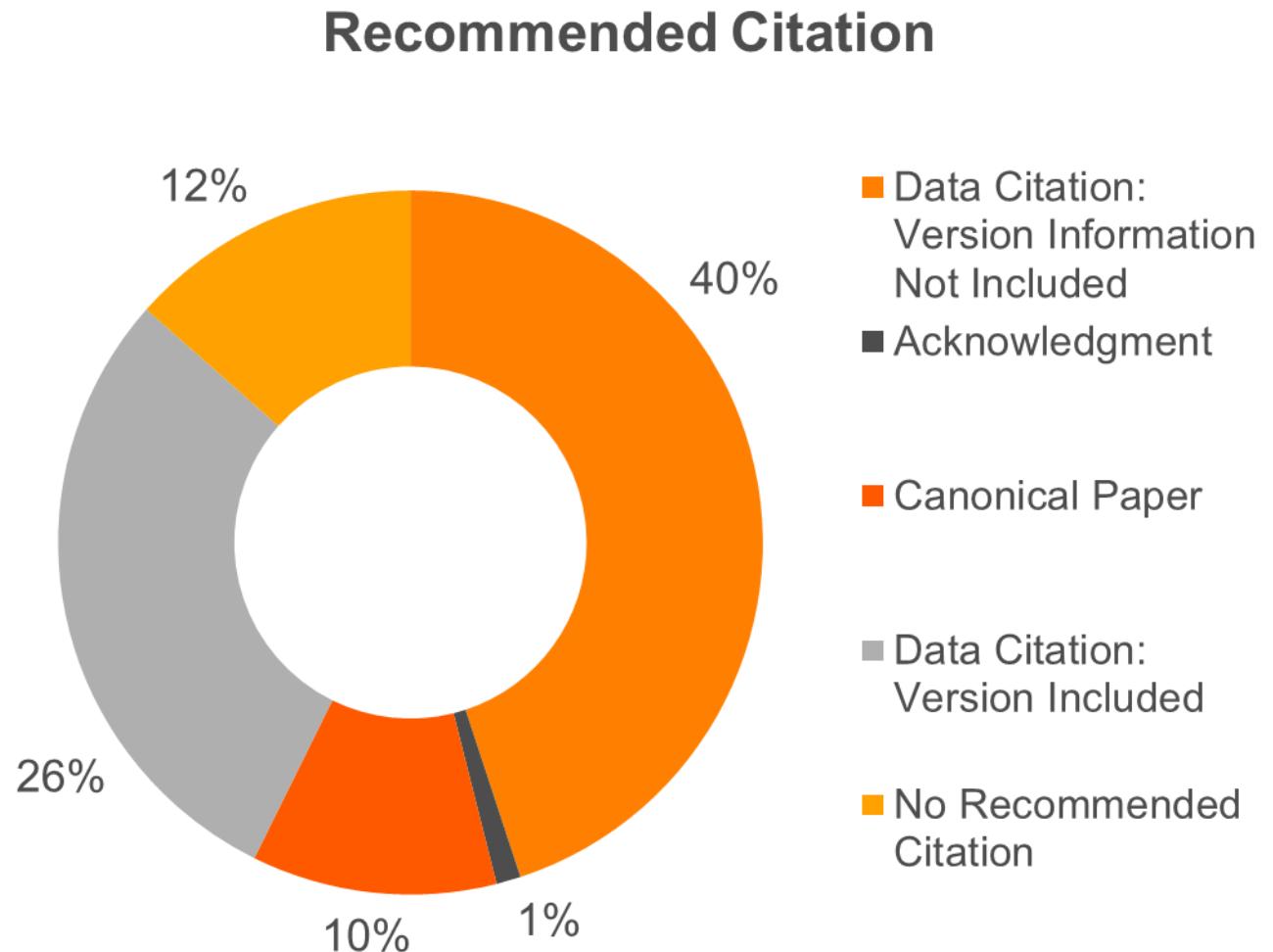
Where is version information found?



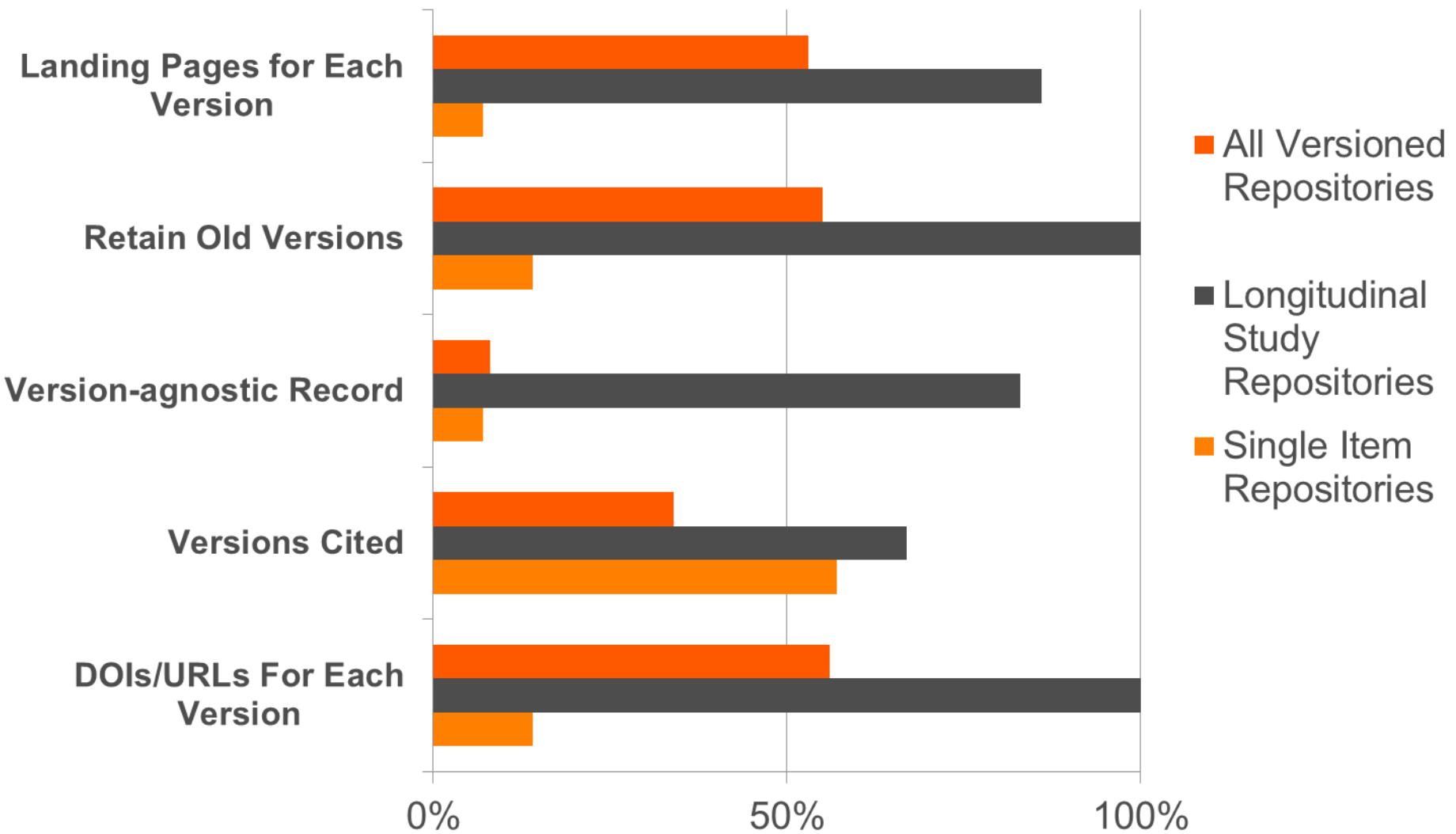
- A significant amount of version information is not readily available through a dedicated metadata tag
- Version information may be concatenated with dataset title, or may be found in the dataset identifier (accession number, DOI, etc)

Versions in citations?

- Only 26% of repositories which employ versioning include version in a recommended data citation
- Versions are included in formal data citations for a greater share of these repositories

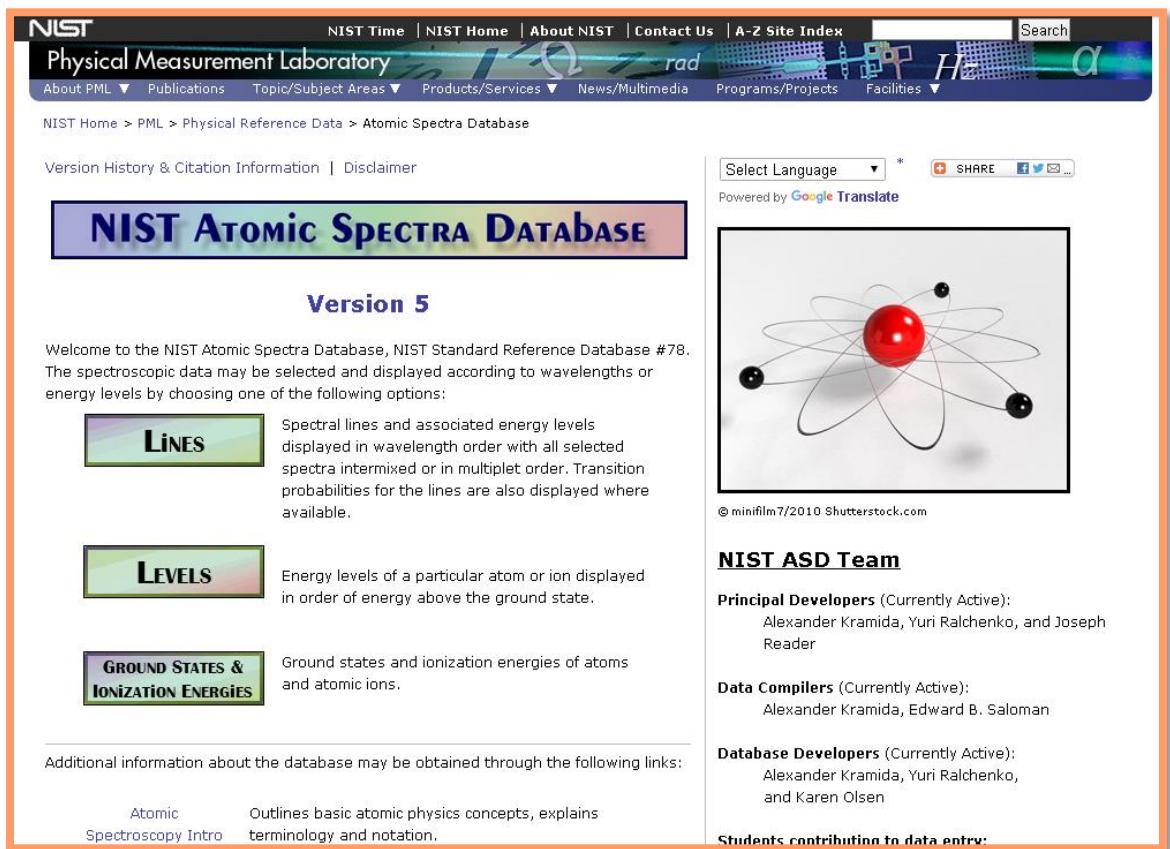


Versioning practices by data repository sub-group



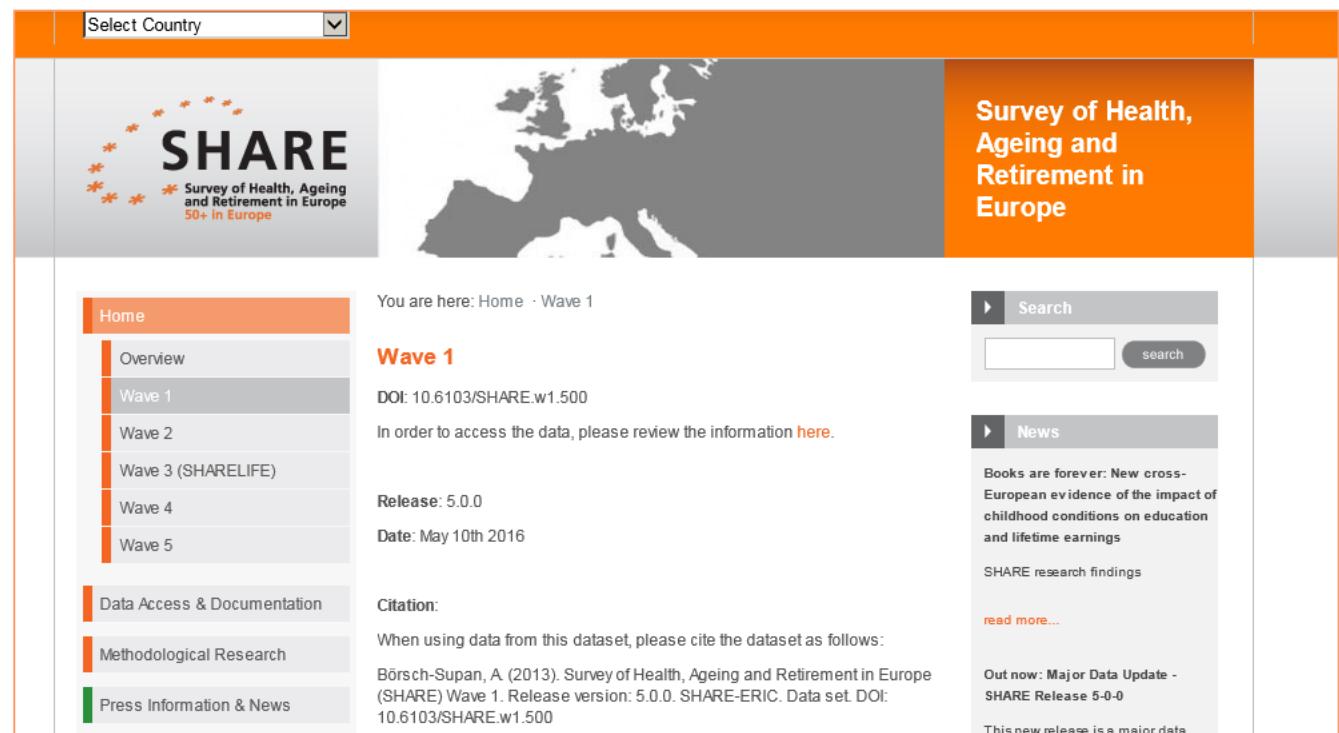
Subgroup: single items

- **DOIs/URLs for each version in 14% of cases**
- **93% do not have a version-agnostic record or landing pages for each version**
- **86% do not retain older versions**
- **For all single items with a recommended data citation, version is included**



Subgroup: longitudinal study repositories

- **67% have version information in title field**
- **100% retain old versions and have DOIs/URLs for each version**
- **83% have version-agnostic record (in this case, the repository record)**
- **86% have landing pages for each version**



The screenshot shows the SHARE website interface. At the top, there is a navigation bar with a dropdown menu for 'Select Country'. The main header features the 'SHARE' logo with the subtitle 'Survey of Health, Ageing and Retirement in Europe' and '50+ in Europe'. To the right of the logo is a map of Europe. The right side of the header is an orange box containing the text 'Survey of Health, Ageing and Retirement in Europe'. The left sidebar has a 'Home' button (highlighted in orange) and links for 'Overview', 'Wave 1' (highlighted in grey), 'Wave 2', 'Wave 3 (SHARELIFE)', 'Wave 4', 'Wave 5', 'Data Access & Documentation', 'Methodological Research', and 'Press Information & News'. The main content area shows the 'Wave 1' landing page. It includes the text 'You are here: Home · Wave 1', the title 'Wave 1', the DOI '10.6103/SHARE.w1.500', and a note: 'In order to access the data, please review the information [here](#)'. It also displays the 'Release: 5.0.0' and 'Date: May 10th 2016'. Below this is a 'Citation' section with the text: 'When using data from this dataset, please cite the dataset as follows: Börsch-Supan, A. (2013). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 5.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w1.500'. To the right, there is a 'Search' bar and a 'News' section with a link to 'SHARE research findings' and a 'read more...' button. At the bottom right, there is a link to 'Out now: Major Data Update - SHARE Release 5-0-0' and a note: 'This new release is a major data'.

Takeaways

- For 19% of repositories which employ versioning, some results were inconclusive, due to all dataset versions being listed as '1.0'
- Repository policies with respect to versioning are not displayed on database websites
- Metadata vs. data versioning is generally unclear/difficult to determine
- Versioning practices may vary significantly even within the datasets of a single data repository
- While versioning practices are being adopted by repositories in the interest of correct citation and reproducibility, little or no guidance exists regarding best practice at a cross-disciplinary or discipline-specific level

Questions/Comments

“Technical mechanisms for citation are only surface characteristics of the knowledge structures in which they are embedded”

- Christine Borgman

Big Data, Little Data, No Data: Scholarship in the Networked World (2015)