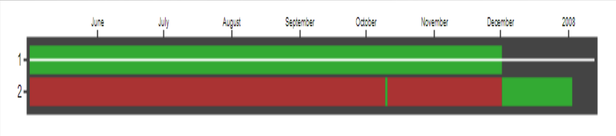
# What is a data version?

Matt Macduff, Sherman Beus

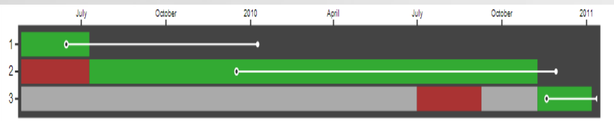
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

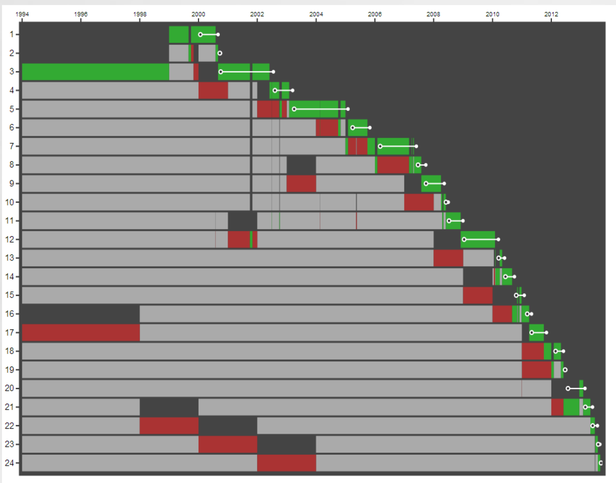
- Reprocessing creates new copies of data.
- Data set changes can be organized into “versions”.
- Version numbers provide a place to link to change documents.
- ARM is considering adding a version attribute to the data.
- Examples of possible data versions:



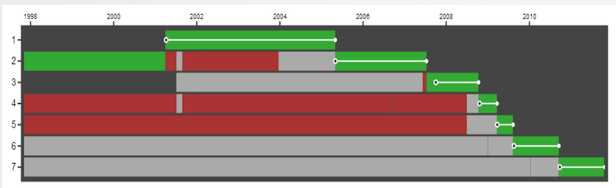
The fkbwacrM1.b1 data was reprocessed near the end of the deployment. This is represented as a new version (2). The data added after the reprocessing is only part of this new version. The visualization shows the first copy of data as green and later copies of the same day as red.



The grwmwrpM1.b1 shows a reprocessing event early on. Near the end, a 2 month period was reprocessed but the rest of the data wasn’t changed. The unchanged data is part of versions 2 and 3. The grey color indicates that the data was not changed. However, the entire data set is only fully available in version 3.



The sgp15okmX1.b1 data has annual reprocessing events for the previous year, but there are several other events for older data. The majority of data in each version is also in many other versions (gray color) but almost all of it has been reprocessed at least once. Versions 22, 23 and 24 happened in rapid succession and were likely just one reprocessing task. The rules for creating new versions could combine this into a single version change. Similarly, versions 13 & 14 or versions 20 & 21 could possibly combined. Reducing the number of versions created makes it easier to document and understand the changes that occurred. This plot also shows an artifact of reprocessing. Prior to storing new data sets, the previous copy is deprecated and no longer available. How this is best captured in version changes is still a work in progress.



The sgpmfrsrE6.b1 originally started in 2001 but later the older ‘a1’ data was reprocessed to produce the b1 data for historical data. We note that version 3 only has a few recent days of data reprocessed. Version 3 could be combined with version 2 but could create confusion for users of the latest version of data. To avoid this confusion, we are considering delaying the assignment of versions to new data. Thus near-term data might have a “pending” or “not assigned” version. While clearly versions 4 and 5 represent major reprocessing tasks, versions 6 and 7 are both related to very limited amount of data. A versioning number scheme that includes major and minor numbers is being considered for these cases. Thus the numbers might instead be 1, 2, 2.1, 3, 4, 4.1, and 4.2.

Data versions provide users a handle to reference and explore the historical changes to data.