

Data Preservation for Re-Use: from tens of TB to tens of EB

This talk discusses the challenges in storing, sharing and curating data from High Energy Physics experiments at CERN and elsewhere. Data volumes range from tens of TB to tens of EB and the target period for which the data should remain fully usable runs to a few decades. (For example, data from LEP, where the active data taking period ran from 1989 to 2000, is expected to be fully usable until around 2030 with the bits available for much longer).

The talk differentiates between what is needed to preserve the data (and the necessary “knowledge” so that it remains usable) from the infrastructure and services required to share the data with future users – sometimes for previously unknown purposes. It also discusses how we measure if we are achieving our goals, including the use of Active Data Management Plans, Certification and agreed metrics for knowledge capture and analysis reproducibility.

To ease comparisons with other disciplines and projects, the talk will loosely follow the [guidelines for Data Management Plans in EU Horizon 2020](#) projects.