Geophysical Research Abstracts
Vol. 14, EGU2012-12278, 2012
EGU General Assembly 2012

Data Reuse and Transparency in the Data Lifecycle

S. Worley and D. Schuster
NCAR, Boulder, United States (worley@ucar.edu, 303-497-1248)

Data reuse and transparency are significantly more important now than in the past. Policy makers, community leaders, independent citizens, and commercial interests all need improved access to reference datasets. Scientific data centers can support the best fact-based outcomes, in traditional research and these expanding communities, by following tested archiving and access principles, and by monitoring and recording system usage.

Data authenticity relies on proven curation practices, throughout the data lifecycle, that guarantee reproducibility of results. These practices include ensuring data integrity and complete archives, preserving all dataset versions with documented lineage, managing diverse data formats, harvesting standardized metadata, tracking user interactions, and having disaster recovery workflows for all data, metadata, documentation, and software. Additionally, stronger linkages between source data and scholarly work, and vice versa, publications to data are needed.

Sustained collaboration with the user community and other data centers is essential for facilitating data reuse. Immutable reference datasets complemented by community derived secondary datasets build a foundation for widespread reuse. As part of the data discovery process, all users should be offered a starting point that is suited to their needs and could be anywhere in the chain of related datasets. At a minimum data centers could acquire related datasets or create links between distributed data, but a more robust and scalable approach is needed. Future solutions will rely on accepting common metadata standards that can be leveraged to build and maintain relationships between local and distributed datasets.

This presentation will expand on the principles of data reuse and reproducibility, in general, and in the context of achievements and goals for the Research Data Archive (RDA) at NCAR. The RDA is a science data resource for weather and climate research that has been developed over 40+ years.