# Data Reuse and Transparency in the Data Lifecycle

Steven Worley
Doug Schuster
Bob Dattore
National Center for Atmospheric Research
Boulder, CO USA

CISL Computational & Information Systems Laboratory

NCAR

# Topics

Data Reuse and Transparency

- What are these data features?
- Why are they important?
- Archiving practices
- Access practices

# What are these data features?

- Data reuse implies:
  - Expanding usage beyond intended primary community
  - Maintaining reference datasets and building many products from them
- Data transparency implies:
  - Reproducibility - ability to reproduce data files or products for users
  - Traceability – tagging and preserving access details

# Why are Reuse and Transparency Important?

Data centers/providers are expected to support fact-based outcomes:

- Traditionally for science/research
- **Now also for policy makers, community leaders, individual citizens, and commercial interests.**

# Supporting New Reuse and Transparency

- Decisions by policy makers
  - Traceable open access sources
- Actions by community leaders
  - Planning for societal services
    - Emergencies, water, energy, etc.
- Usage by citizens and educators
  - Inquisitive science, family activities, safety
  - Science learning
- Collaborative commercial applications
  - Tighter coupling between engineering and science
    - Wx forecasts for wind energy production
    - Energy companies contribute mesoscale observations

# Archiving practices

- Curation that assures data authenticity
  - Preserve original data formats, to the max. extent possible.
    - Maintaining 100% content and accuracy – serious challenge
- Use a "rich" metadata standard
  - A local standard?
  - Generate discipline and cross-discipline standards
    - E.g. ISO, DIF, etc.
- Create multiple copies
  - Data files, metadata, documentation, and software
  - Disaster recovery – not a secondary concern

# Archiving practices

- Collection completeness and integrity
  - Closely monitor data work flow
    - Account for every file
  - Read every file
    - Gather, check, preserve metadata
  - Compute and preserve file checksums
- Maintain dataset lineage / provenance
  - Use approved processes to delete datasets (never?)
  - Establish tiered "level of service" for data
    - Move old / superseded versions to lower level
  - Keep all metadata on the highest tier – discoverable!

# Archiving practices

- Explicit data version tracking
  - Sometimes, internal to files
  - Always, within data management system
  - Include notations in all documentation
- Establish Digital Object Identifiers (DOIs)
  - Two-way linkage between publications and data
  - Promotes easy path for follow-on research from publications
  - Leverages skills / facilities of libraries – richer knowledge base
  - Create data family tree connections

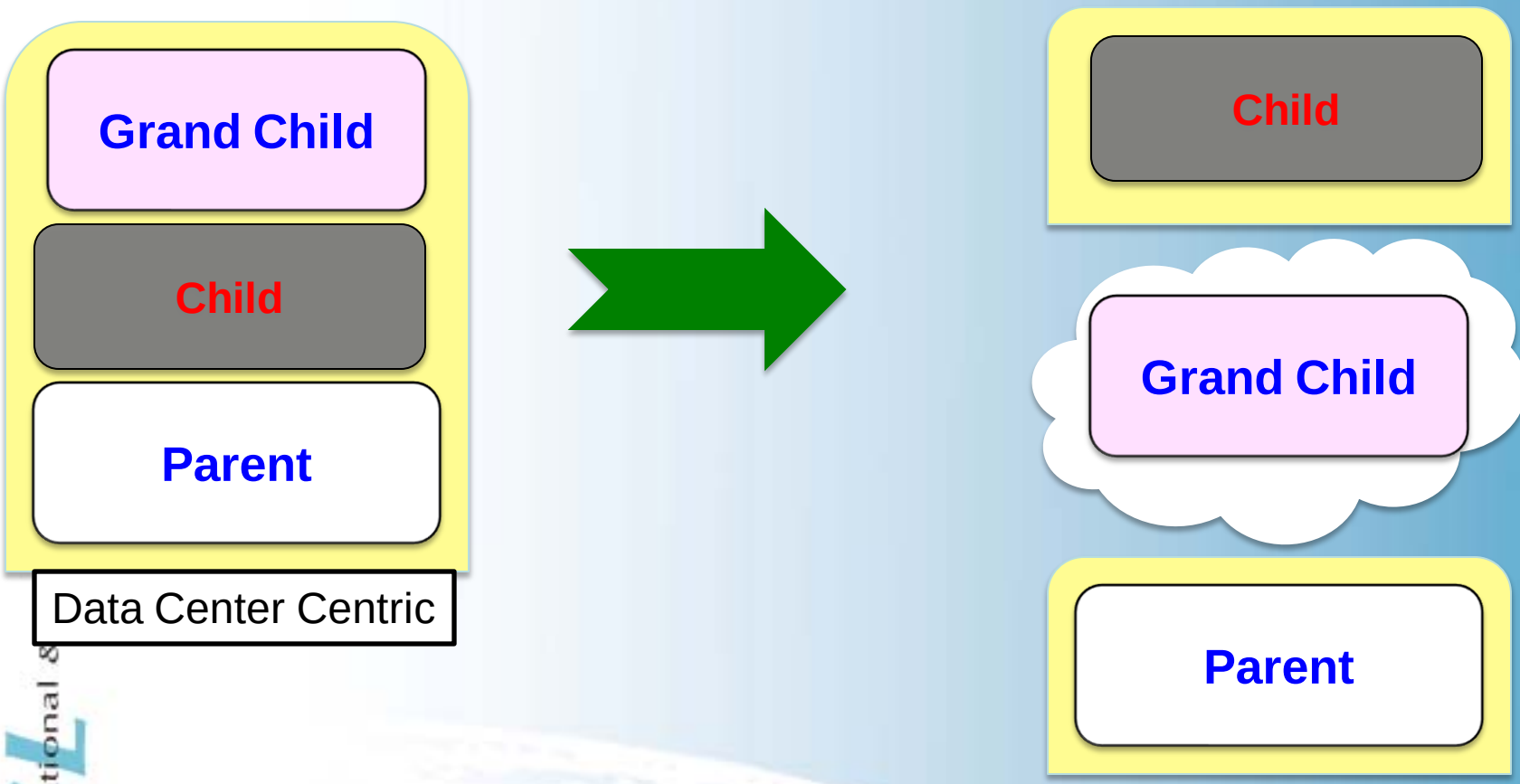
# Dataset Family Tree Example

**Global and Regional Atmospheric and Ocean Re-analyses**
NCEP/NCAR, NARR, ERA-40, ERA-Interim, 20CR, OARCA

| | | |
|---|---|---|
| **NOC Surf. Flux** (1973-2009) | **WASwind** (1950-2009) | **Etc.** |
| **Ocean Clouds** (1900-2010) | **JMA SST** (1871-2011) | **HadSLP** (1871-2011) |
| **HadISST** (1871-2011) | **NOAA OI SST** (1981-2011) | **NOAA ERSST** (1854-2011) |

**International Comprehensive Ocean Atmosphere Data Set (ICOADS)**
Global marine surface observations (1662-2011)

# Dataset Family Tree - Evolution




**Challenges:**

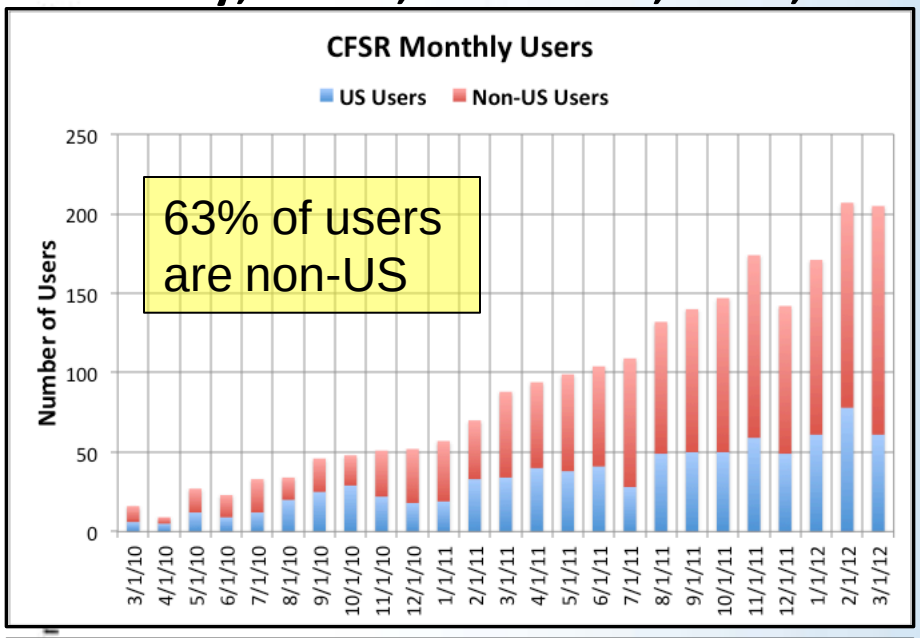
- **System of immutable IDs – DOIs?**
- **Multi-institution preservation commitment**
- **Transparency across institutions, accepted standards/governance**
- **Promote discovery by sharing metadata, OAI-PMH**
- **Future, knowledge-based discovery and access via ontologies within semantic web**

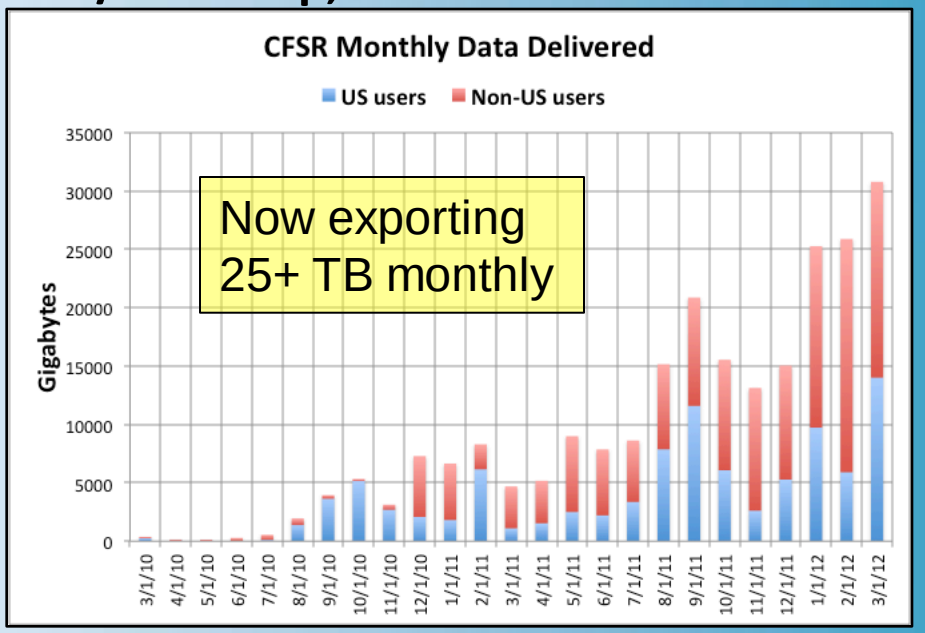
# Access Practices

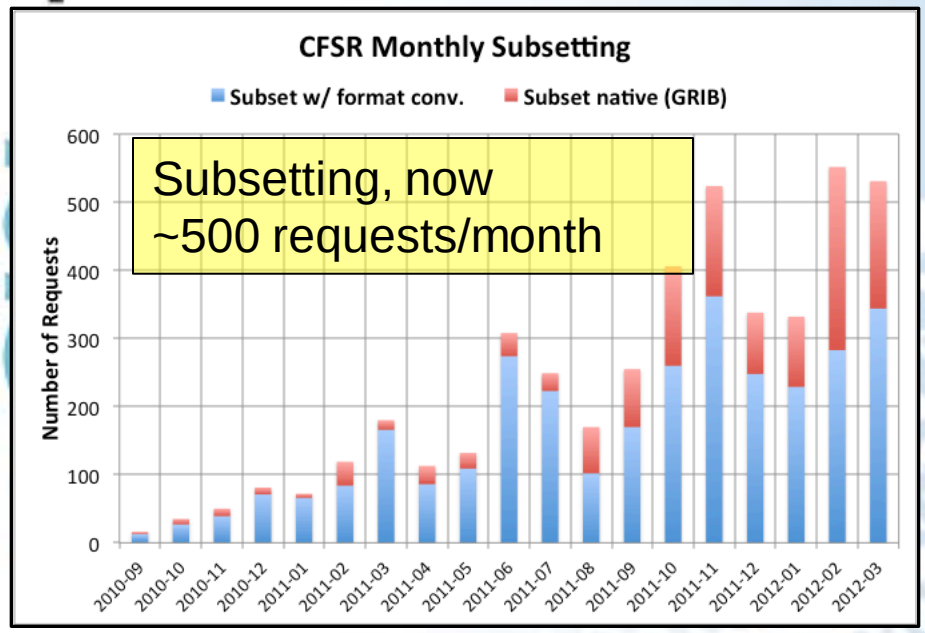
- User Identification – key to reproducibility
- Record all data access transactions
  - Who received what and when
  - Log product creation constraints from GUIs and web services
  - Log software IDs used for product creation
- Benefits
  - Reproduce a data access process
  - Feedback to users about data changes
  - Use metrics imply how to improve access

# Metrics Example

## CFSR 6hrly, GRIB2, 1979-2011, 75TB, 28K fields/time step, 168K files




Track User activity:
- who accessed what and when

# Conclusions

- Data reuse and transparency are rapidly expanding in importance
- Many "best practices" in archive management support reuse and transparency
- Archive access monitoring is necessary for transparency, reproducibility, and traceability
- Need significant improvement in linking data family trees and data to publications to advance reuse and transparency

# Research Data Archive at NCAR
# http://rda.ucar.edu/