Tracking data usage at NCAR's Research Data Archive

Steven Worley

Computational and Information System Laboratory

NCAR

CISL Computational & Information Systems Laboratory

NCAR

Topics

- Current practices @ NCAR's Research Data Archive
- Data citations with or without unique identifiers
- Tracking challenges not addressed by citations - **alone**

Current Practices @ NCAR's Research Data Archive

Knowing who the users are permits data usage tracking

- Initiated user registration in 2005
  - Online, automatic, only verify the email address
  - Self-declare five registration fields
  - Manage information in MySQL DB
  - Not shared, changeable, recovery capable
  - **All data downloads are traceable to a user**

Current Practices @ NCAR's Research Data Archive

Computational & Information Systems Laboratory

CISL

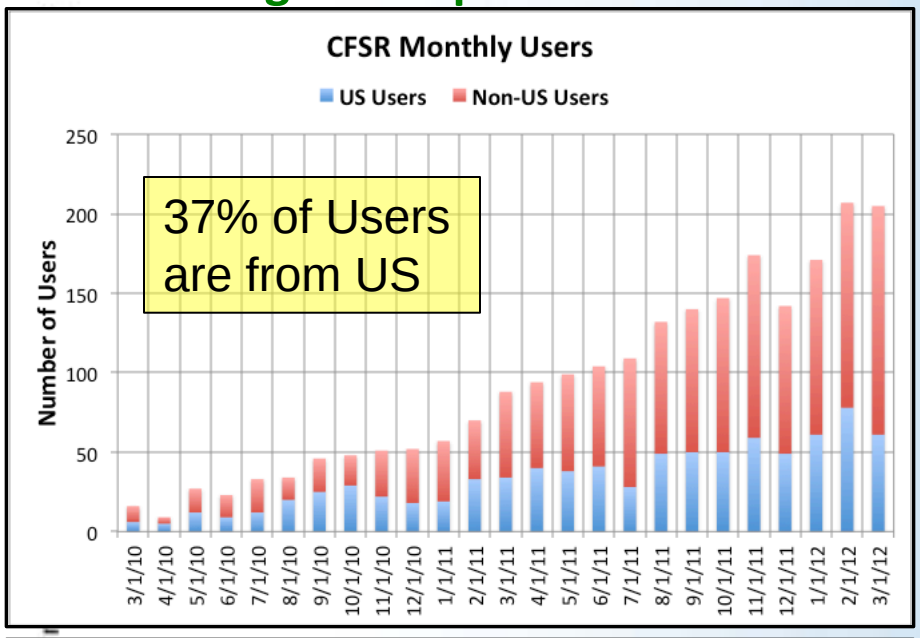
NCAR

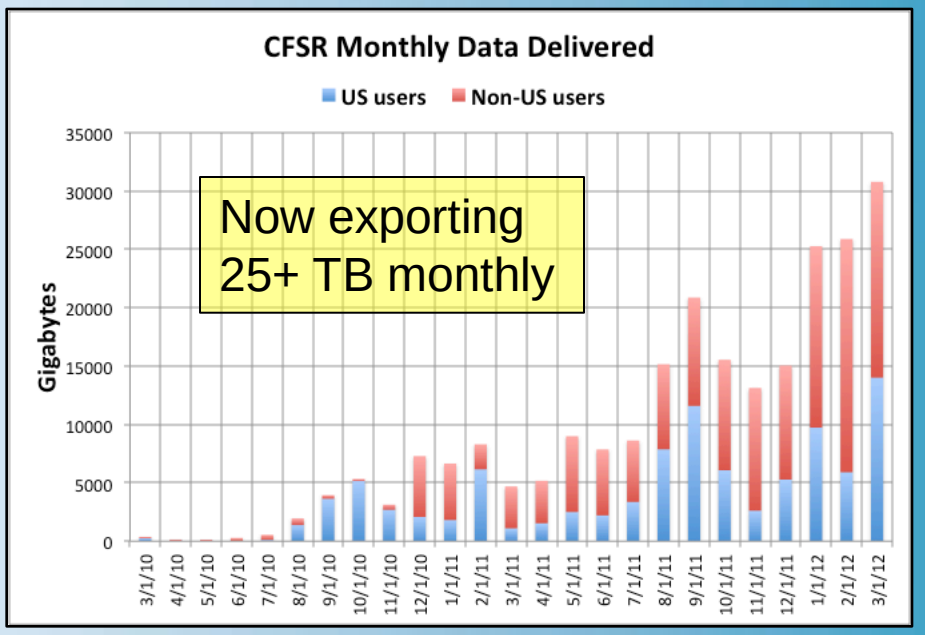
Can analyze metrics in many ways

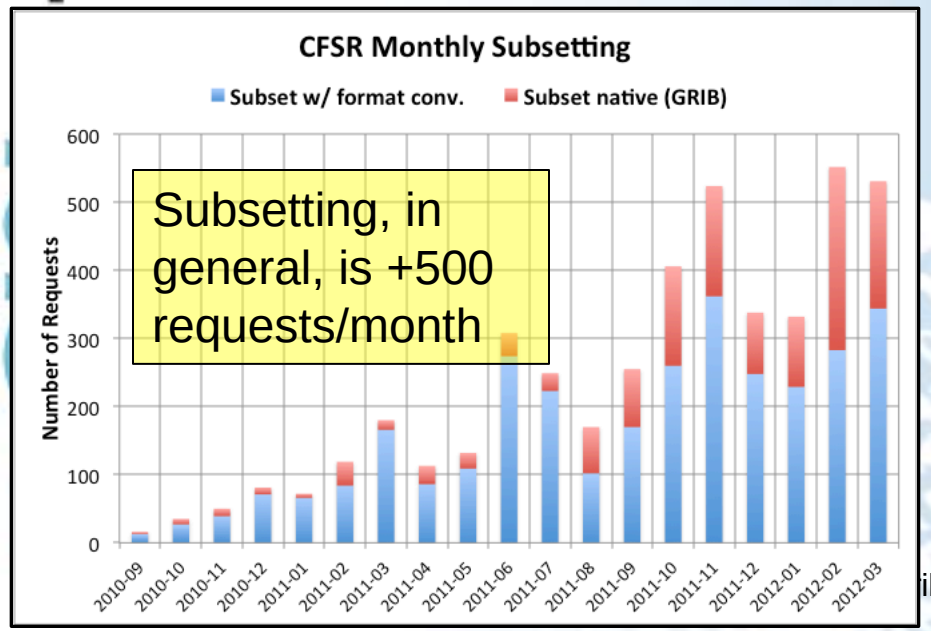
	Email addr.	Name	Org. Type	Org. Name	Org. Country
Single file	☑	☑	☑	☑	☑
Dataset	☑	☑	☑	☑	☑
Access date	☑	☑	☑	☑	☑
Service type	☑	☑	☑	☑	☑
Data amount	☑	☑	☑	☑	☑
Format conversion	☑	☑	☑	☑	☑
Subset constraints	☑	☑	☑	☑	☑

Current Practices @ NCAR's Research Data Archive

Metrics Usage - Sample







Track User activity:
- who accessed what and when

Current Practices @ NCAR's Research Data Archive

User Perspectives

- One time registration – not synchronized with larger schemes (e.g. OpenID)
- Does not inhibit scripted access or web services
- Safeguard against erroneous data
  - Current example: JRA-25 is under repair
  - 1900 files out of +1M need to be replaced
  - Spread across 22 data products
  - **Notify, via email, all users who accessed erroneous files**

Data Citations with or without unique identifiers

Without Identifiers

- RDA has a recommended acknowledgement text with each dataset
- No easy way to quantify scientific impact – ineffective

With Identifiers

- DOIs would build immutable links between publications and data
- Implementation – previously discussed by Mayernik, Daniels, Strand, and Kaplan

Tracking challenges not addressed by citations - alone

Strengthening DOI Movement

- Make data citation with DOI ubiquitous in publications
- Get helpful push from funding agencies
  - Review DMPs for data publication including DOIs
  - 30 March; Dear Colleague Letter – Data Citation in the Geosciences
    - http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp?WT.mc_id=USNSF_25&WT.mc_ev=click
- Need institutions to recognize data publication as a career-worthy contribution

CISL Computational & Information Systems Laboratory

Tracking challenges not addressed by citations - alone

Taking Advantage of the DOI Movement

- Develop tools to evaluate data DOIs in publications
  - Determine data service impact on science
  - Measure the rate of change of scientific productivity?
- Develop the data to publications linkage
  - Reverse engineer publication citations back to datasets
  - Create dataset citation bibliographies – provide forward links to publications during data discovery

CISL Computational & Information Systems Laboratory

Tracking challenges not addressed by citations - alone

Integrating data resources – leveraging DOIs

- Better support research by informing users about related datasets
  - Parent / Child = Raw-data / Derived-data
  - DOIs, because they are unique, are a key asset here

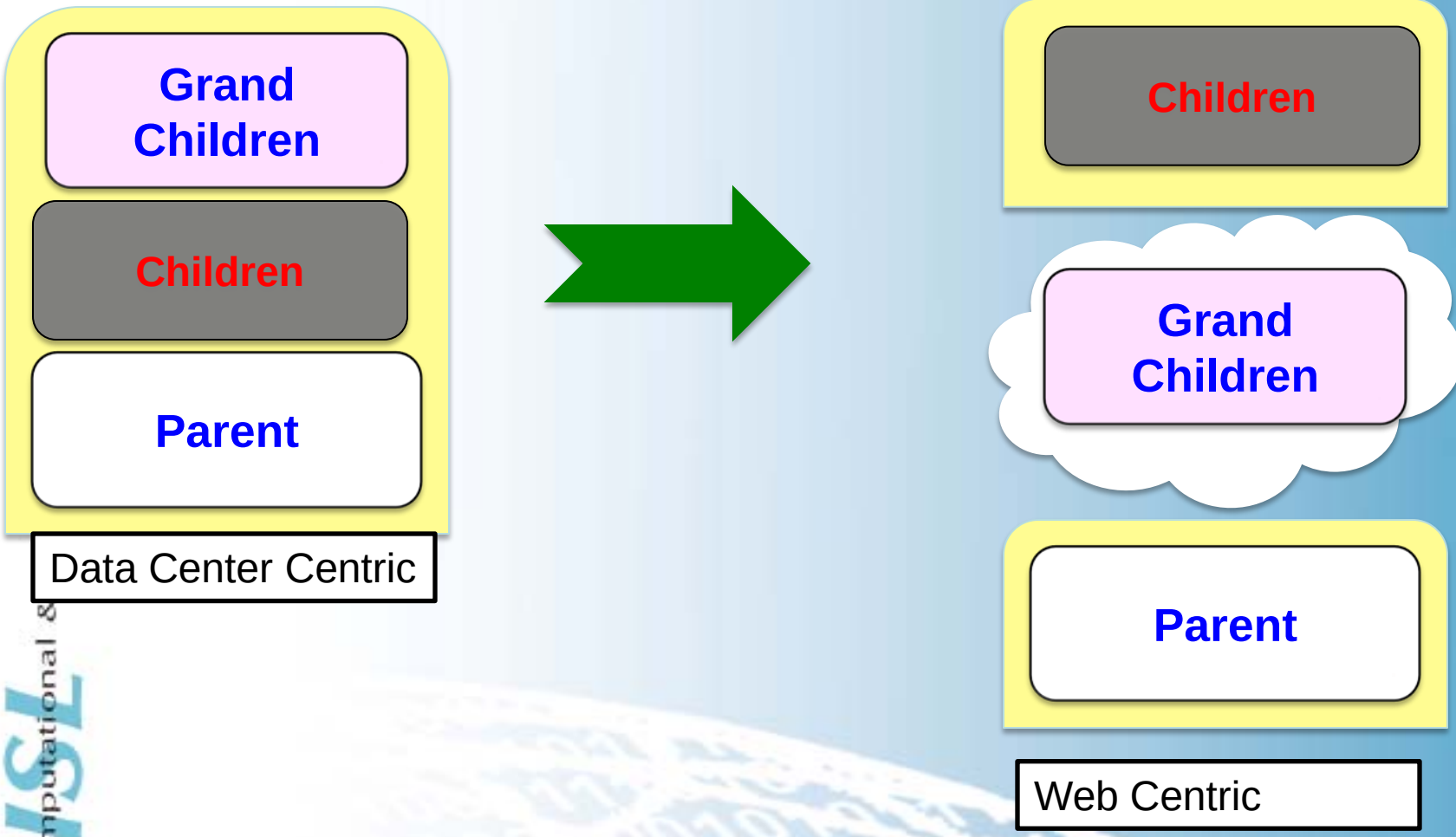
Dataset Family Tree Example

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



How can we provide users with dataset relationships in a distributed environment?

Solution Proposal:

- Create system-wide standard metadata
- Create capacity at data DOI repositories to accept "dataset relationship metadata"
- Publish and share in a standard form
  - OAI-PMH and ontologies
- Exploit the inherent "meaning" in the relationships

Example: standard metadata & capacity at data DOI repositories

DataCite has established a starting point

- Optional DataCite Property – **RelatedIdentifier**
  - Attribute **relatedIdentfierType** (DOI, ARK, URL,…)
  - Attribute **relationType** (Compiles, IsVariantFormOf…)
- Possible need **relationType**,(DerivesTo, IsDerivedFrom)?
- Meaningful metadata connecting datasets
  - <relatedIdentifier relatedIdentifierType="DOI" relationType="IsDerivedFrom">10.1234/nnnnnn</relatedIdentifier>

Publish and share in a standard form

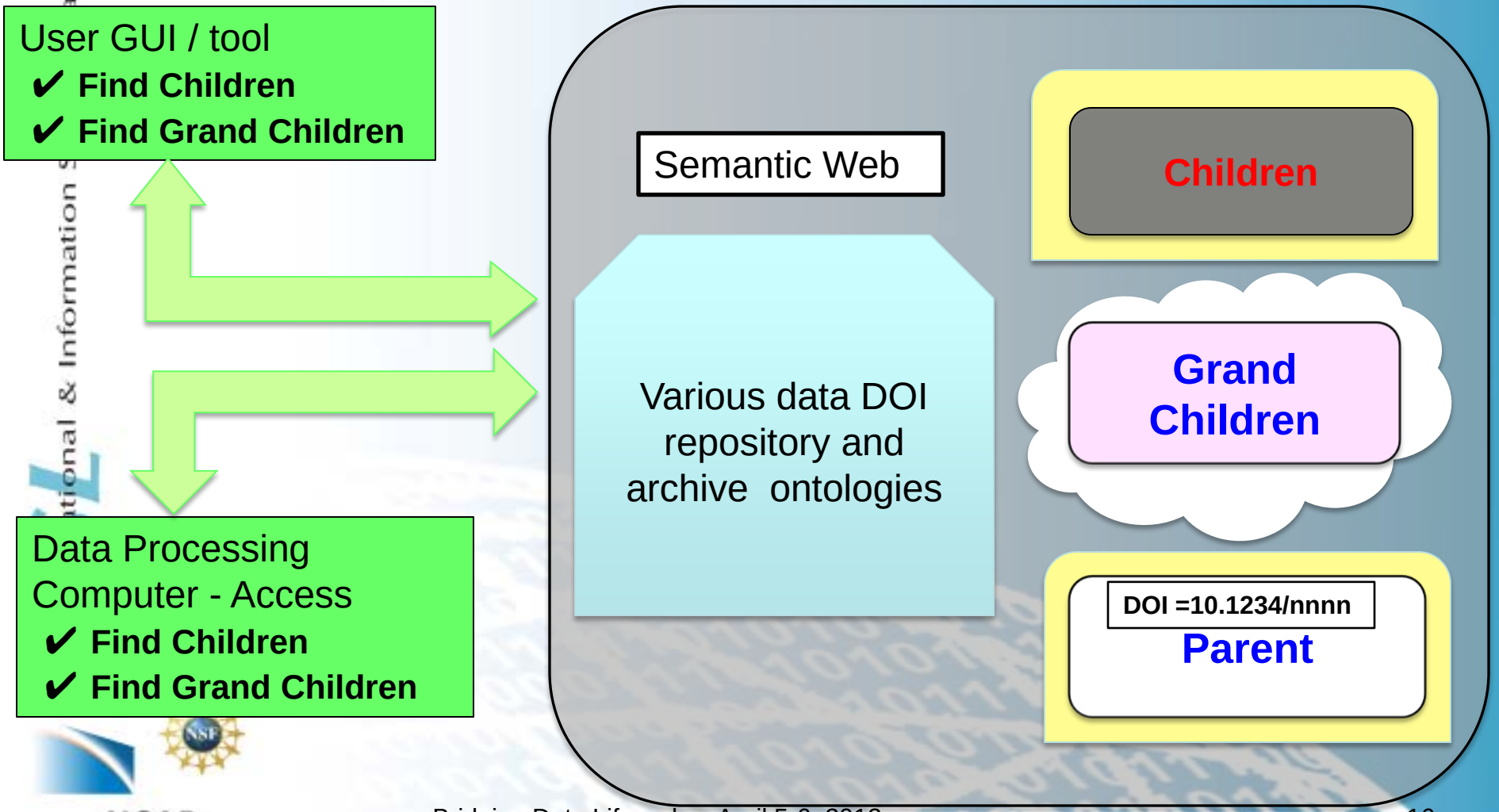
Two bulk exchange methods

- OAI-PMH Service
  - Open Archives Initiative Protocol for Metadata Harvesting
  - Open access to individuals
- Metadata schemes can be mapped to ontologies
  - Casts metadata into a “standard” semantic reference framework (e.g. RDF and publish it)
  - **DataCite Ontology** (http://purl.org/spar/datacite/)
  - Ontologies enable:
    - Metadata to be interrogated with programming logic
    - Automatic integration with similar data

CISL Computational & Information Systems Laboratory

Knowledge-based question:

"What datasets are derived from DOI = 10.1234/nnnn?"



Summary

- Current practices @ NCAR's Research Data Archive
- Data citations with or without unique identifiers
- Tracking challenges not addressed by citations - **alone**

Steven Worley
worley@ucar.edu

The Questions

- *How do data archive/repositories currently track data use?*
- *How would data citations (with or without unique web identifiers) help to make tracking data use easier?*
- *Are there challenges related to tracking data uses that data citations do not help to address?*