

Managing Dataset DOIs and Versions in a Changing Archive

Steven Worley

Bob Dattore

Zaihua Ji

National Center for Atmospheric Research
Boulder, Colorado, USA

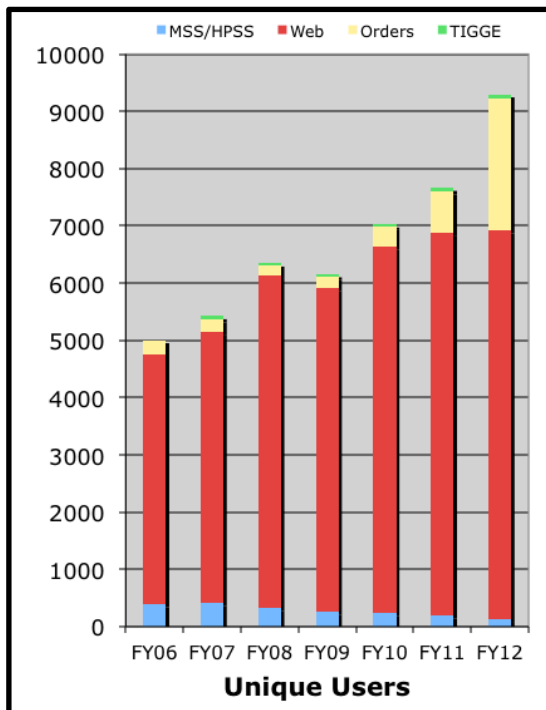
Topics

- RDA Background
- Use Cases
- User DOI Services



Research Data Archive (RDA) at NCAR

1. 600+ distinct datasets for climate and weather research, 8M Files
2. Collections: ocean & atmosphere observations, analyses, reanalyses, operational NWP outputs
3. Free and open access



CISL Research Data Archive
Managed by NCAR's Data Support Section
 Data for Atmospheric and Geosciences Research

RDA

<http://rda.ucar.edu>

Go to Dataset:

Home
Find Data
Ancillary Services
About/Contact
For Staff

Look For Data:

All Datasets	Variable/Parameter	Type of Data
Time Resolution	Platform	Spatial Resolution
Topic/Subtopic	Project/Experiment	Supports Project
Data Format	Location	Recently Added/Updated

Other Ways to Explore:

- **GCMD Topic:**
[Agriculture](#) | [Atmosphere](#) | [Biosphere](#) | [Climate Indicators](#) | [Cryosphere](#) | [Hydrosphere](#) | [Land Surface](#) | [Oceans](#) | [Paleoclimate](#) | [Solid Earth](#) | [Spectral/Engineering](#) | [Sun-earth Interactions](#)
- **Reanalyses:**
[All Reanalysis Datasets](#) | [ECMWF ERA15 Reanalysis \(ERA15\)](#) | [ECMWF ERA40 Reanalysis Project \(ERA40\)](#) | [ECMWF Interim Reanalysis \(ERA-I\)](#) | [Japanese 25-year Reanalysis \(JRA25\)](#) | [NCEP Climate Forecast System Reanalysis \(CFSR\)](#) | [NCEP/NCAR Reanalysis Project \(NNRP\)](#) | [NOAA-CIRES 20th Century Reanalysis \(20CR\)](#)
- **Station Observations:**
[Land Surface Air Temperature: Hourly, Monthly](#)

Dataset Search:

Keyword(s) [Advanced Options](#)

Get Help:

- [Frequently Asked Questions](#)
- [Reset your password](#)
- [A-Z Site Index](#)
- [RDA Users Email List](#)

Recent News:

New Domain Name for the CISL RDA Web Server
 At 1:30pm MDT on Thursday, April 5, the CISL RDA web server will begin operating ...

Spatial Subsetting Now Available for Select Gridded Datasets
 Users now have the option to get spatial area data subsets from select gridded RDA ...

Precipitation validation data now available from the TIGGE subset data portal
 Users requesting model validation data in addition to model forecast data from the TIGGE data ...

[Archive...](#)

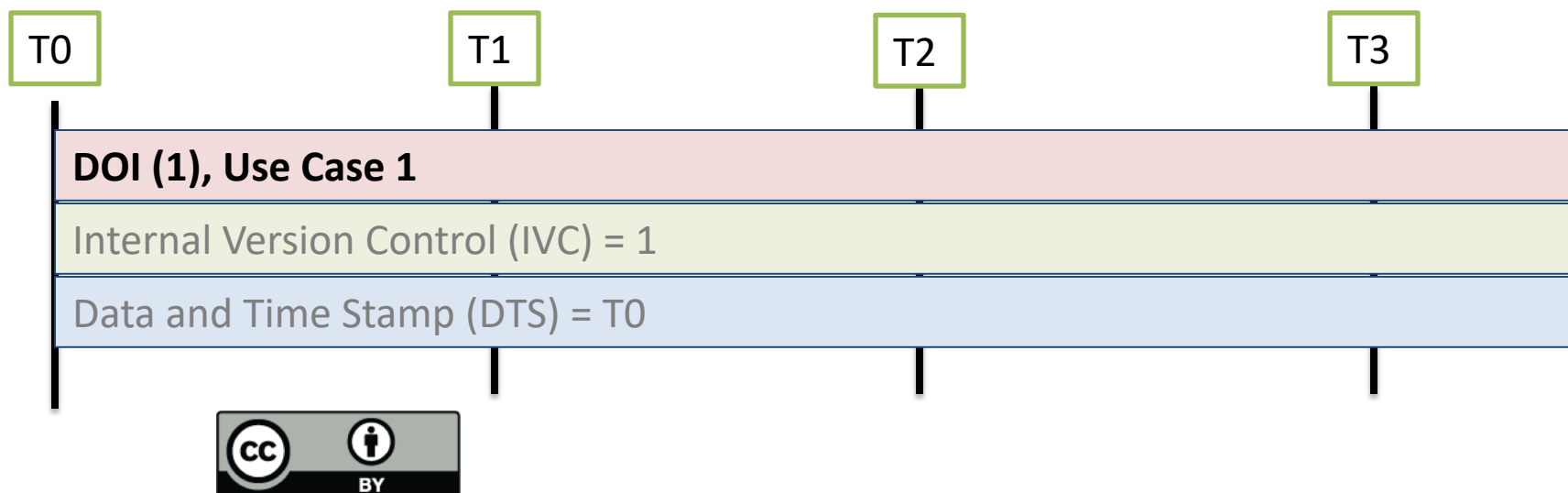
Technical Approach – MySQL DB

- DB records for each file
 - DOI
 - Internal Version Control (IVC) setting
 - Date and Time Stamp (DTS) of file activities
- Other features
 - Maintain file to dataset relationship
 - Maintain history of file activities
 - Tracks user access via registration and login



Use Case 1 – Create a new DOI dataset

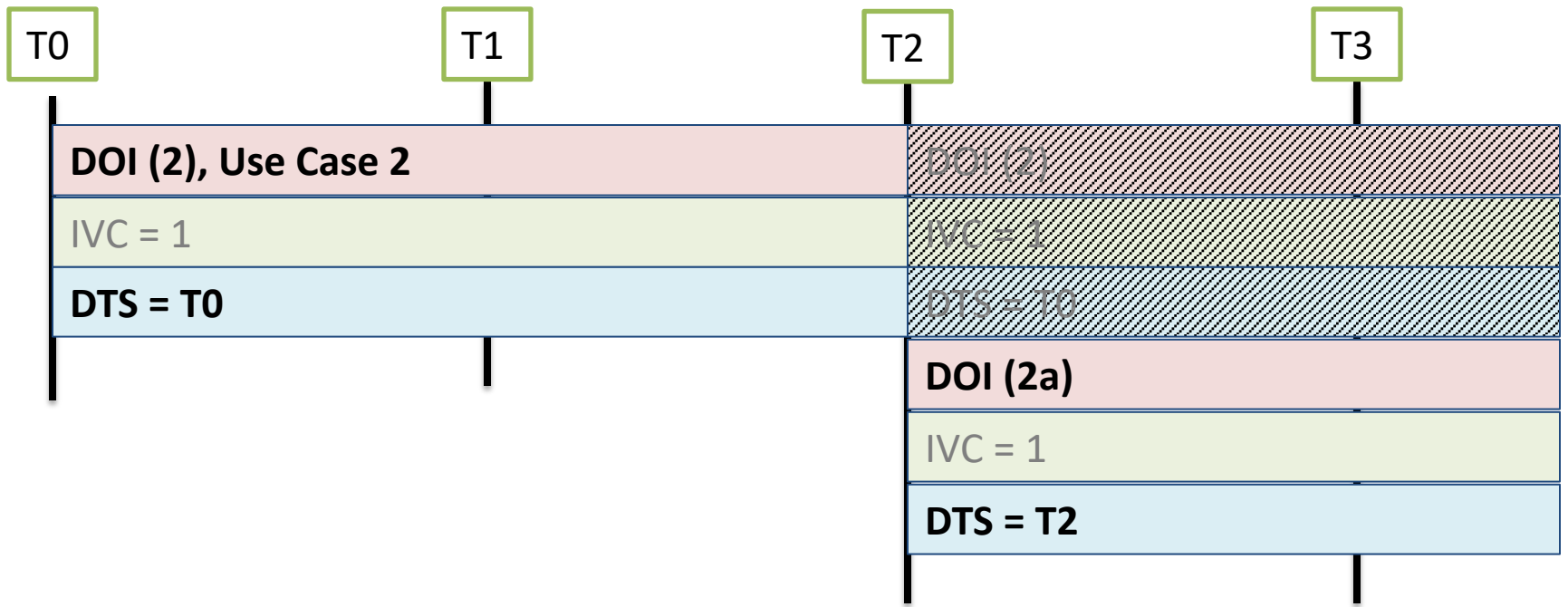
- All files one-to-one match on tape (offline) and online storage
 - Exceptions: permit Endian byte swap, standard file packaging (tar, gzip, htar, etc.)
- Mint a new DOI through DataCite



Use Case 2 – Complete dataset replacement (e.g. new data from the provider)

- RDA dataset landing page (URL) is unchanged
 - Metadata (discovery, file content) updated
- Assign new DOI
- Old version
 - Files offline – tape archive
 - File-set permanently frozen
 - Create new landing page (URL) for old DOI
 - Inform user of options
 - Go to new DOI
 - Initiate recovery of old DOI file-set
 - Update the URL in DataCite metadata

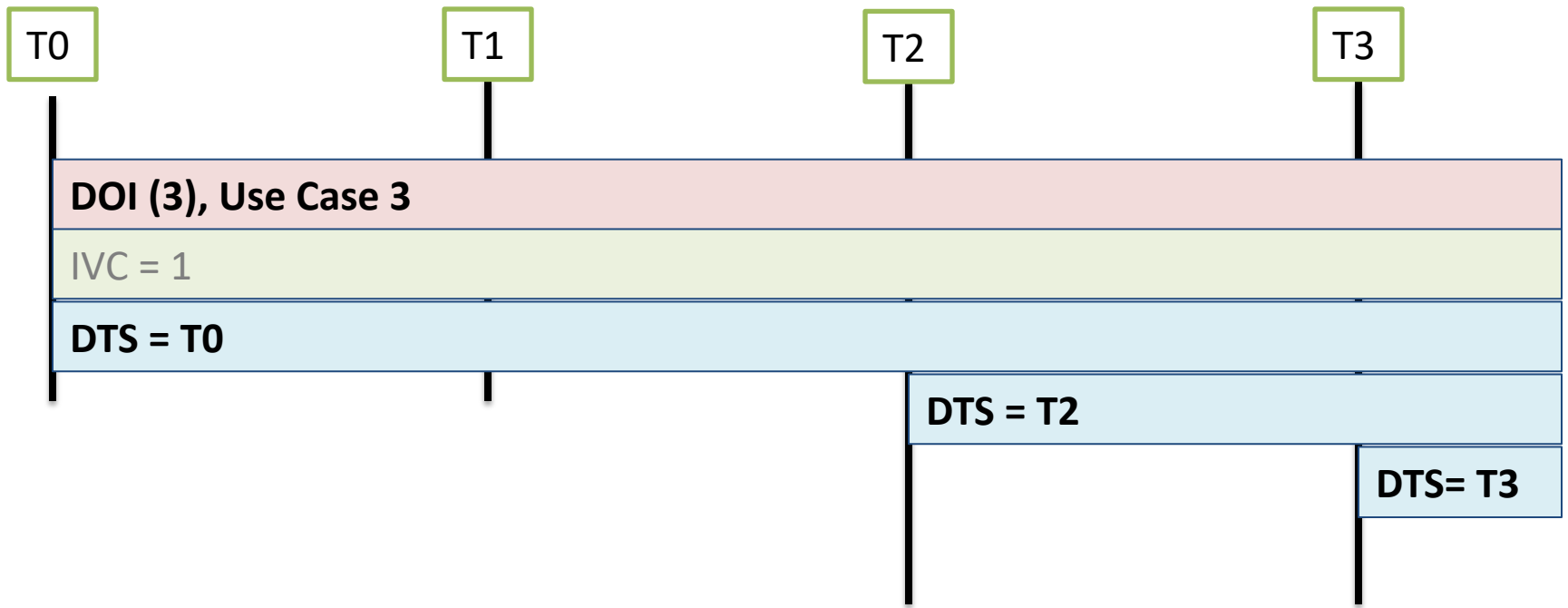




Use Case 3 – Routine dataset extension in time

- Add new files
 - Inherit existing DOI and IVC
 - Log DTS into DB
 - Allow adding data to the newest file
 - E.G. Adding monthly data to an annual file
 - Update DTS
 - Data replacement is not permitted
- Regularly update temporal coverage in DataCite metadata
 - Frequency: monthly or weekly (TBD)

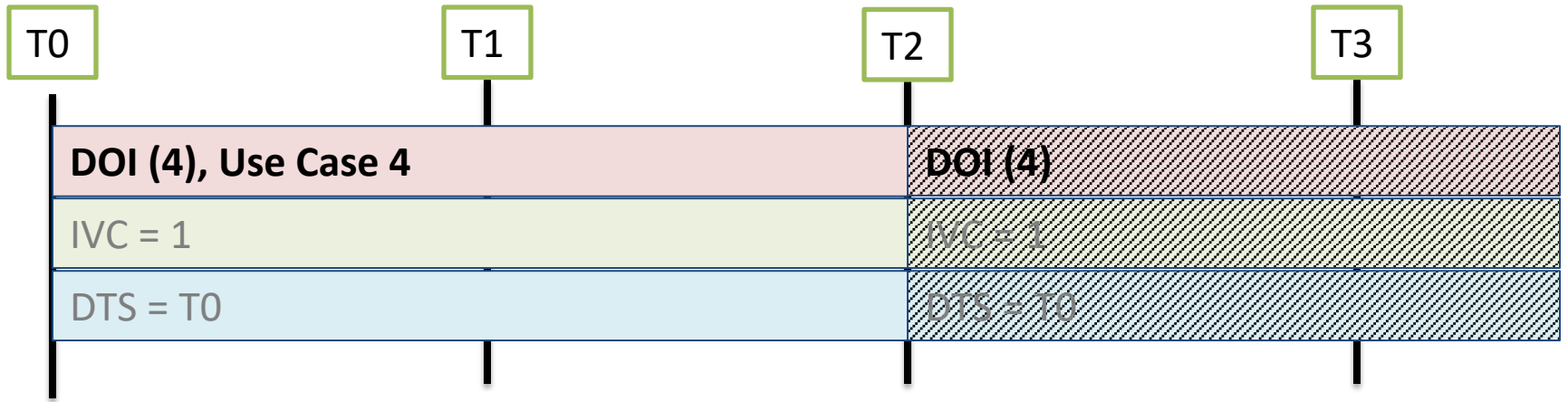




Use Case 4 – Removal of a DOI dataset

- Update DataCite metadata so DOI resolves to a special “dead” dataset landing page (URL)
- Landing page explains status and options
 1. File set is preserved and can be restaged
 - Use Case 2, recover from tape (offline) archive
 2. File set has been deleted from the system
 - Explanation required





Use Case 5 – Small scale replacement (fixes) within a dataset

- Erroneous files are removed from the file-set
 - Files permanently preserved
 - IVC and DTS are saved as history in DB
- Actions to replace a file
 - Incremented IVC, $nn \rightarrow nn+1$
 - Re-assign IVC across complete file set
 - Add IVC notation to replacement file base-name
 - » noaa_CFR_hourly_1988_2mTemp_**IVC2**.grb
- DOI remains unchanged



T0

T1

T2

T3

DOI (5), Use Case 5

IVC = 1

DTS = T0

File Replacement
F1-9 @T1 (orange) F1-9 @T0 (hatched)

IVC = 2

DTS = T0 (left) DTS = T1 (right, red border)

File Replacement
F1-9 @T1 (orange) F100-120 @T2 (orange) F100-120 @T0 (hatched)

IVC = 3

DTS = T0 (left) DTS = T1 (middle, red border) DTS = T2 (right, red border)



User DOI services

Citation design – ESIP Guidelines

Compo, G. P., et al. 2010. *International Surface Pressure Databank (ISPDv2) 1768 to 2010. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory.*
<http://dx.doi.org/10.5065/D6SQ8XDW>. Accessed[§] dd mmm
YYYY.

[§] Please fill in the “Accessed” date with the day, month and year (e.g. – 5 Aug 2011) you last accessed the data from the RDA.

Also offer AMS, AGU, DataCite styles as an option.



Download standard metadata for citation management software, e.g. Endnote, Zotero, etc.



User DOI services

Three ways to get a citation

1. Generic dataset citation, from RDA portal
2. Download service (scripts, subsetting): Provide complete dataset citation, including “Accessed on” date.
3. Generate citations on demand at a later time:
 - Display user specific access activities
 - Utilize registration information
 - Allow activity selection
 - Create the complete citation



Some Outstanding Challenges

- No limit on data sharing after extraction from the RDA
 - Could lose ability to provide accurate citations
- Have not designed a way to tag an access event with the software ID used to enable it
 - E.g. format conversion, regridding, server-side computations
- Have not designed a systematic way to couple DOIs from the RDA with nearly identical or related datasets
 - Could be managed with metadata enhancements



Conclusions

- Managing DOIs for a dynamic archive has complications
 - Full dataset replacements
 - Dataset retirements
 - Routine dataset extension
 - Stewardship improvements – data fixes, patches, etc
- Implementation – keep records for each file, including:
 - DOI
 - Internal Version Control
 - Date and Time Stamp
- Provide users options to create citations, base on ESIP recommendations



Questions?

RDA: <http://rda.ucar.edu>

DataCite: <http://www.datacite.org>

ESIP:

http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations

(Federation of Earth Science Information Partners)

