

Data management services supported by NCAR's Research Data Archive

Thomas Cram and Doug Schuster

National Center for Atmospheric Research

Computational and Information Systems Laboratory (CISL)

AMS 99th Annual Meeting

Paper TJ 22.4
10 January 2019

About the RDA

- **History**
 - Established 1960s
- **Purpose**
 - Support climate & weather research at NCAR and UCAR universities with reference datasets
- **Collections**
 - Ocean & atmospheric observations, climate reanalyses, operational NWP products
 - 600+ datasets, 10M files, 2.2 PB
 - Continually growing: 70+ updated daily-monthly
- **Free and open access**
- **Science educated staff**



rda.ucar.edu



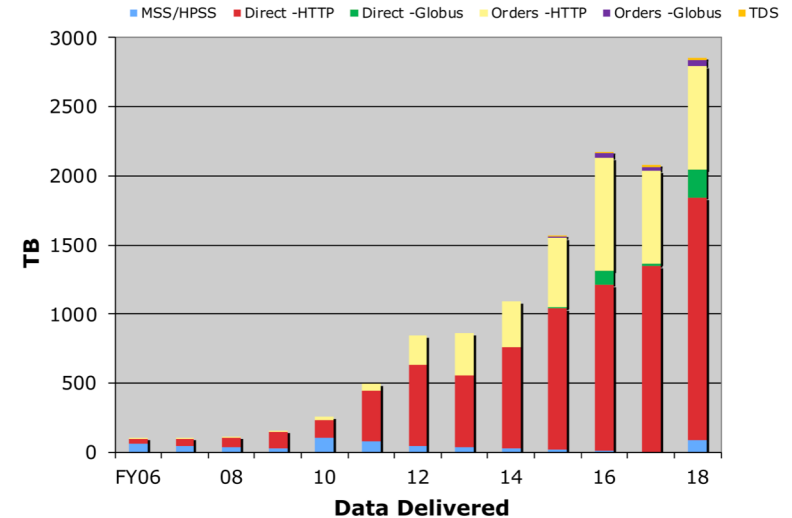
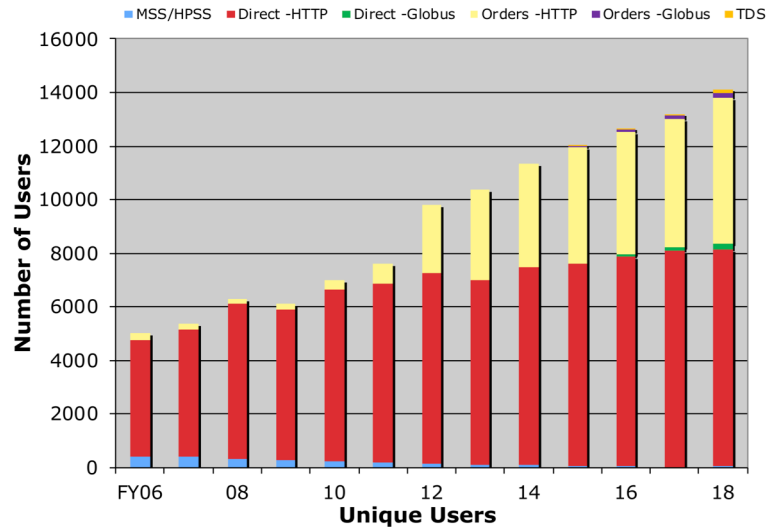
RDA Data usage



- **FY 2018**

14K+ unique web users

2.85 PB data delivered



RDA user services

- Challenges & strategies



- **Enable reproducible research**
 - Structured DOI policies
 - User history of data access maintained to facilitate dynamic citation generation
- **Make data discovery easy & relevant**
 - Maintain robust metadata DB
- **Provide scalable data access capabilities**
 - Diversity of data formats/dialects
 - Growth in user numbers and types
 - Growth in archive volume and complexity
- **Goal: Reduce time spent dealing with data**

The screenshot displays the RDA web interface. At the top, there are "Download Options" with instructions on how to download files. Below this is a table titled "Internet Download Files - Create a File" with columns for Filename, Data Format, File Contents, Valid Date Range, Size, Subset ID, and More Details. The table lists several files related to the GPCP Version 2.3 Monthly Analysis Product. Below the table, there is a "Dataset Product" section with a map of the United States and a "Your Access History for September 2018" calendar. At the bottom, there is a "Dataset Citation" section with a citation for the GPCP Version 2.3 Monthly Analysis Product.

Download Options:

- Select one or more files and download them using [Globe](#) [What is Globe?](#)
- Select two or more files and [Download](#) them as a single Unix tar file
- Select one or more files and [Create](#) [a](#) [Unix](#) [script](#) to read them from the Internet using [wget](#)
- Click the individual filename links to download files one-at-a-time

Reset checkboxes | Range selection is off | on | Total volume of 5 selected files: 110.69 Mbytes

Filename	Data Format	File Contents	Valid Date Range	Size	Subset ID	More Details
CMORPH_V0_x_RAW_0.25deg-3HLY_20170201	NCEP CPC CMORPH025	8 Grids	2017-02-01 00:00	22.12	V0.x.0.25DEG-2017	More Details
CMORPH_V0_x_RAW_0.25deg-3HLY_20170202	NCEP CPC CMORPH025	8 Grids	2017-02-02 00:00	22.12	V0.x.0.25DEG-2017	More Details
CMORPH_V0_x_RAW_0.25deg-3HLY_20170203	NCEP CPC CMORPH025	8 Grids	2017-02-03 00:00	22.12	V0.x.0.25DEG-2017	More Details
CMORPH_V0_x_RAW_0.25deg-3HLY_20170204	NCEP CPC CMORPH025	8 Grids	2017-02-04 00:00	22.12	V0.x.0.25DEG-2017	More Details
CMORPH_V0_x_RAW_0.25deg-3HLY_20170205	NCEP CPC CMORPH025	8 Grids	2017-02-05 00:00	22.12	V0.x.0.25DEG-2017	More Details

Dataset Product: Unrestricted GDAS
Temporal Selection: 2017-02-01 to 2017-02-05
Platform Type: Land_station
Spatial Selection: [for detailed info] Enter a full or partial platform ID: - OR - Using the interactive map and the observations (the default is the entire globe)

Dataset: GPCP Version 2.3 Monthly Analysis Product® (ds728.4)
Your Access History for September 2018: Choose a day to get a citation for this dataset. You will also see more details about your downloads on that day, which will help you verify that this is the citation you want.

September 2018

Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

For Data Accessed on 2018-09-25:

Dataset Citation: [RIS](#)

Mesoscale Atmospheric Processes Branch/Laboratory for Atmospheres/Earth Sciences Division/Science and Exploration Directorate/Goddard Space Flight Center/NASA, and Earth System Science Interdisciplinary Center/University of Maryland. 2018, updated monthly. GPCP Version 2.3 Monthly Analysis Product. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. <https://doi.org/10.5065/D6SN07QX>. Accessed 25 Sep 2018.

Bibliographic citation shown in: Federation of Earth Science Information Partners (ESIP) | style

Data Access Detail:

Data citation – Technical approach

- Registration through DataCite.org
- Collection level DOI, i.e. one DOI per data collection
- MySQL DB records for each file
 - DOI
 - Internal Version Control (IVC)
 - Date and time stamp of file activities
- Other features
 - Maintain history of file activities
 - Tracks user access via registration and login



The screenshot shows the NCAR Research Data Archive website. The header includes the NCAR logo, the text 'Research Data Archive Computational & Information Systems Lab', and 'NCAR is sponsored by National Science Foundation'. A navigation bar has links for Home, Find Data, Ancillary Services, About/Contact, Data Citation, and Web Services. The main content area is for the dataset 'JRA-55: Japanese 55-year Reanalysis, Daily 3-Hourly and 6-Hourly Data' with DOI 10.5065/D6HH6H41. It includes a description, data access information, and a list of publications citing the dataset. An abstract is also provided.

How to Cite This Dataset:

RIS

BibTeX

Japan Meteorological Agency/Japan. 2013, updated monthly. *JRA-55: Japanese 55-year Reanalysis, Daily 3-Hourly and 6-Hourly Data*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. <https://doi.org/10.5065/D6HH6H41>. 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. Bibliographic citation shown in | Federation of Earth Science Information Partners (ESIP) | style

[Get a customized data citation](#)



User DOI services

Three ways to get a citation

1. Generic citation, from RDA portal
 - Multiple citation formats

How to Cite This Dataset:
RIS
BibTeX

Compo, G. P., et al. 2015. *The International Surface Pressure Databank version 3*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory.
<https://doi.org/10.5065/D6D50K29>. 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.
Bibliographic citation shown in style

[Get a customized data citation](#)

User DOI services

Three ways to get a citation

1. Generic citation, from RDA portal
2. Download service (scripts, subsetting):
Provide complete dataset citation,
including "Accessed on" date

How to Cite This Dataset:

[RIS](#)

[BibTeX](#)

Compo, G. P., et al. 2015. *The International Surface Pressure Databank version 3*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory.

<https://doi.org/10.5065/D6D50K29>. 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.

Bibliographic citation shown in [Federation of Earth Science Information Partners \(ESIP\)](#) Style

[Get a customized data citation](#)

User DOI services

Three ways to get a citation

1. Generic 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

Get a Dataset Citation

This interface will allow you to view your data download history from the RDA and generate an appropriate data citation for specific data that you have downloaded and wish to cite.

Since the date that the data were accessed is a key component of the data citation, please choose the latest date for a dataset that you have accessed over a range of time.

How to use this tool:

- **By Dataset:** If you know the RDA dataset that you want to cite, use this option to see a list of all datasets from which you have downloaded data.
- **By Date:** If you don't remember the dataset but you remember approximately when you downloaded the data, use this option to see all of the times that you have downloaded data from the RDA. This will help you find the dataset that you want to cite.

By Dataset **By Date**

<< Back to dataset list

The following table shows the months in which you downloaded data from **GPCP Version 1.2 One-Degree Daily Precipitation Data Set (ds728.3)** (ds728.3). Choose a month to see a detailed download history and get a data citation.

Filter by:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016												
2017												
2018												

Dataset:

GPCP Version 1.2 One-Degree Daily Precipitation Data Set[®] (ds728.3)

Your Access History for October 2018:

Choose a day to get a citation for this dataset. You will also see more details about your downloads on that day, which will help you verify that this is the citation you want.

October 2018						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

For Data Accessed on 2018-10-15:

Dataset Citation: **RIS**

Huffman, G. J., D. T. Bolvin, and R. F. Adler. 2016. *GPCP Version 1.2 One-Degree Daily Precipitation Data Set*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. <https://doi.org/10.5065/D6D50K46>. Accessed 15 Oct 2018.

Bibliographic citation shown in style

Data Access Detail:

1 file downloaded from the RDA web server:
http://rda.ucar.edu/data/ds728.3/p1d/netcdf/gpcp_1dd_v1.2_p1d.199705.nc

Dataset citation tracker

- Citations retrieved via
 - CrossRef.org API (journal articles)
 - Google API (books, chapters, etc.)

The screenshot displays the NCAR Research Data Archive website for the ERA-Interim Project. The header includes the NCAR and UCAR logos, the Research Data Archive name, and the Computational & Information Systems Lab affiliation. It also notes that NCAR is sponsored by the National Science Foundation. Navigation tabs include Home, Find Data, Ancillary Services, About/Contact, Data Citation, and Web Services. The dataset title is 'ERA-Interim Project' with the identifier 'ds627.0' and DOI '10.5065/D6CR5RD9'. A search bar is visible with the text 'Go to Dataset:'. Below the title, there is a section for 'Help with this page' and 'Data Citations'. The 'Data Citations' section lists 13 publications that have cited the dataset, starting from 2018. The 'Abstract' section at the bottom states: 'ERA-Interim represents a major undertaking by ECMWF (European Centre for Medium-Range Weather Forecasts) to produce a reanalysis with an improved atmospheric model and assimilation system which replaces'.

NCAR
UCAR

Research Data Archive
Computational & Information Systems Lab

NCAR is sponsored by
National Science Foundation

Go to Dataset:

For assistance, contact Dave Stepaniak (303-497-1343).

ERA-Interim Project
ds627.0 | DOI: 10.5065/D6CR5RD9

Description Data Access Documentation Software

Help with this page: RDA dataset description page video tour

Data Citations: This dataset has been cited 13 times.
* Published works that cited this dataset:

2018

Chandramouli, K., and C. Balaji, 2018: Ingesting microwave sounder radiances for improvement in track forecast of cyclone Vardah. *J. Appl. Rem. Sens.*, **12**, 1, <https://doi.org/10.1117/1.JRS.12.026015>

Huang, J., W. Tian, L. J. Gray, J. Zhang, Y. Li, J. Luo, and H. Tian, 2018: Preconditioning of Arctic Stratospheric Polar Vortex Shift Events. *J. Climate*, **31**, 5417-5436, <https://doi.org/10.1175/JCLI-D-17-0695.1>

Jayasankar, C. B., K. Rajendran, and S. Surendran, 2018: Monsoon Climate Change Projection for the Orographic West Coast of India Using High-Resolution Nested Dynamical Downscaling Model. *J. Geophys. Res. Atmos.*, <https://doi.org/10.1029/2018JD028677>

Khanna, J., D. Medvigy, G. Fisch, and T. T. de Araujo Tiburtino Neves, 2018: Regional Hydroclimatic Variability Due To Contemporary Deforestation in Southern Amazonia and Associated Boundary Layer Characteristics. *J. Geophys. Res. Atmos.*, **123**, 3993-4014, <https://doi.org/10.1002/2017JD027886>

Kutka, E., J. Hubbard, T. Eichler, and A. Lupo, 2018: Symmetry of Energy Divergence Anomalies Associated with the El Niño-Southern Oscillation. *Atmosphere*, **9**, 342, <https://doi.org/10.3390/atmos9090342>

Lowman, L., T. Wei, and A. Barros, 2018: Rainfall Variability, Wetland Persistence, and Water-Carbon Cycle Coupling in the Upper Zambezi River Basin in Southern Africa. *Remote Sensing*, **10**, 692, <https://doi.org/10.3390/rs10050692>

Markert, K. N., R. E. Griffin, A. S. Limaye, and R. T. McNider, 2018: "Spatial Modeling of Land Cover/Land Use Change and its Effects on Hydrology Within the Lower Mekong Basin", in Land-Atmospheric Research Applications in South and Southeast Asia. Ed. K. P. Vadrevu, T. Ohara, and C. Justice, Springer, 667-698.

Meng, X., and J. Cheng, 2018: Evaluating Eight Global Reanalysis Products for Atmospheric Correction of Thermal Infrared Sensor—Application to Landsat 8 TIRS10 Data. *Remote Sensing*, **10**, 474, <https://doi.org/10.3390/rs10030474>

2017

Boothe, A. C., and C. R. Homeyer, 2017: Global large-scale stratosphere-troposphere exchange in modern reanalyses. *Atmos. Chem. Phys.*, **17**, 5537-5559, <https://doi.org/10.5194/acp-17-5537-2017>

Lee, H., R. Z. Bar-Or, and C. Wang, 2017: Biomass burning aerosols and the low-visibility events in Southeast Asia. *Atmos. Chem. Phys.*, **17**, 965-980, <https://doi.org/10.5194/acp-17-965-2017>

Rieckh, T., R. Anthes, W. Randel, S. Ho, and U. Foelsche, 2017: Tropospheric dry layers in the tropical western Pacific: comparisons of GPS radio occultation with multiple data sets. *Atmos. Meas. Tech.*, **10**, 1093-1110, <https://doi.org/10.5194/amt-10-1093-2017>

2016

Rodrigues, C. V., J. Palma, N. Vasiljević, M. Courtney, and J. Mann, 2016: Coupled simulations and comparison with multi-lidar measurements of the wind flow over a double-ridge. *J. Phys.: Conf. Ser.*, **753**, 032025, <https://doi.org/10.1088/1742-6596/753/3/032025>

2015

Chudrum, S., E. Burke, R. Essery, J. Bolke, M. Langer, M. Heikenfeld, P. Cox, and P. Friedlingstein, 2015: An improved representation of physical permafrost dynamics in the JULES land-surface model. *Geosci. Model Dev.*, **8**, 1493-1508, <https://doi.org/10.5194/gmd-8-1493-2015>

Abstract: ERA-Interim represents a major undertaking by ECMWF (European Centre for Medium-Range Weather Forecasts) to produce a reanalysis with an improved atmospheric model and assimilation system which replaces

Dataset citation tracker

- Citations retrieved via
 - CrossRef.org API (journal articles)
 - Google API (books, chapters, etc.)
- *Current metrics (Jan. 2019)*
 - 55 journal articles
 - 3 book chapters

The screenshot shows the NCAR Research Data Archive website. At the top, it says "NCAR UCAR | Research Data Archive Computational & Information Systems Lab" and "NCAR is sponsored by National Science Foundation". Below this is a navigation bar with links: Home, Find Data, Ancillary Services, About/Contact, Data Citation, Web Services. A search bar on the right says "Go to Dataset: nna.na". The main content area is for the "ERA-Interim Project" with DOI: 10.5065/D6CR5RD9. It includes a "Description" tab, "Data Access", "Documentation", and "Software" tabs. The "Description" tab is active, showing "Help with this page: RDA dataset description page video tour" and "Data Citations: This dataset has been cited 13 times." Below this is a list of publications citing the dataset, starting with "2018 Chandramouli, K., and C. Balaji, 2018: Ingesting microwave sounder radiances for improvement in track forecast of cyclone Vardah. J. Appl. Rem. Sens., 12, 1, https://doi.org/10.1117/1.JRS.12.026015". The "Abstract:" section at the bottom states: "ERA-Interim represents a major undertaking by ECMWF (European Centre for Medium-Range Weather Forecasts) to produce a reanalysis with an improved atmospheric model and assimilation system which replaces".



Dataset citation tracker

- Citations retrieved via
 - CrossRef.org API (journal articles)
 - Google API (books, chapters, etc.)
- *Current metrics (Jan. 2019)*
 - 55 journal articles
 - 3 book chapters
- Adherence to citation guidelines (authors)
 - Enforce at review stage?
- Dependent on publishers providing info to CrossRef.org

The screenshot shows the NCAR Research Data Archive (RDA) website. At the top, it says "NCAR UCAR | Research Data Archive Computational & Information Systems Lab" and "NCAR is sponsored by National Science Foundation". Below this is a navigation bar with links: Home, Find Data, Ancillary Services, About/Contact, Data Citation, Web Services. A search bar on the right says "Go to Dataset: nna.n". Below the navigation bar is a section for the "ERA-Interim Project" with the DOI "10.5065/D6CR5R09". It includes a description, data access, documentation, and software tabs. The "Data Citations" section lists 13 citations, including works by Chandramouli, K., and C. Balaji; Huang, J., W. Tian, L. J. Gray, J. Zhang, Y. Li, J. Luo, and H. Tian; Jayaraman, C. B., K. Rajendran, and S. Surendran; Khanna, J., D. Medvigy, G. Fisch, and T. T. de Araujo Tiburtino Neves; Kutta, E., J. Hubbart, T. Eichler, and A. Lupo; Lowman, L., T. Wei, and A. Barros; Markert, K. N., R. E. Griffin, A. S. Limaye, and R. T. McNider; Meng, X., and J. Cheng; and Rodrigues, C. V., J. Palma, N. Vasiljević, M. Courtney, and J. Mann. The "Abstract" section states: "ERA-Interim represents a major undertaking by ECMWF (European Centre for Medium-Range Weather Forecasts) to produce a reanalysis with an improved atmospheric model and assimilation system which replaces..."

Metadata processing & support

- Maintain programmatic data ingest capabilities
- Support a range of dataset search/discovery mechanisms
 - Metadata Standards (ISO-19137, NASA-DIF, FGDC, etc.)
 - Controlled vocabularies – NASA GCMD
 - Access points (OAI-PMH, CSW)
 - Schema.org embedded JSON-LD metadata (to support Google Dataset Search)
- Support data subsetting services
 - 97% data download reduction

Google Dataset Search

The screenshot shows the Google Dataset Search interface. The search bar at the top contains the text "NOAA/CIRES Twentieth Century Global Reanalysis Version 2c". Below the search bar, it indicates "7 results found". The first result is highlighted with a grey bubble and shows the UCAR logo, the dataset title "NOAA/CIRES Twentieth Century Global Reanalysis Version 2c", the URL "rda.ucar.edu", and the date "Published Mar 16, 2015". To the right of the search results, there is a detailed view of the first result, showing the UCAR logo, the dataset title, the URL "rda.ucar.edu", the DOI link "https://doi.org/10.5065/D6N877TW", the publication date "Mar 16, 2015", and a list of authors including Gilbert Compo, Linden Ashcroft, Renate Auchmann, Mac Benoy, Pierre Bessemoulin, Theo Brandsma, Philip Brohan, Manola B. Thomas Cram, Richard Crouthamel, Jeffrey Whitaker, Pavel Groisman, Hans Hersbach, Philip Jones, Trausti Jonsson, Sylvie J. Knapp, Andries Kruger, Hisayuki Kubota, Gianluca Lentini, Prashant Sardeshmukh, Andrew Lorrey, Neal Lott, Sandra Lubker, Ju Marshall, Maurizio Mauger, Cary Mock, Hing Mok, Oyvind Nordli, Rajmund Przybylak, Robert Allan, Mark Rodwell, Thomas Ros.

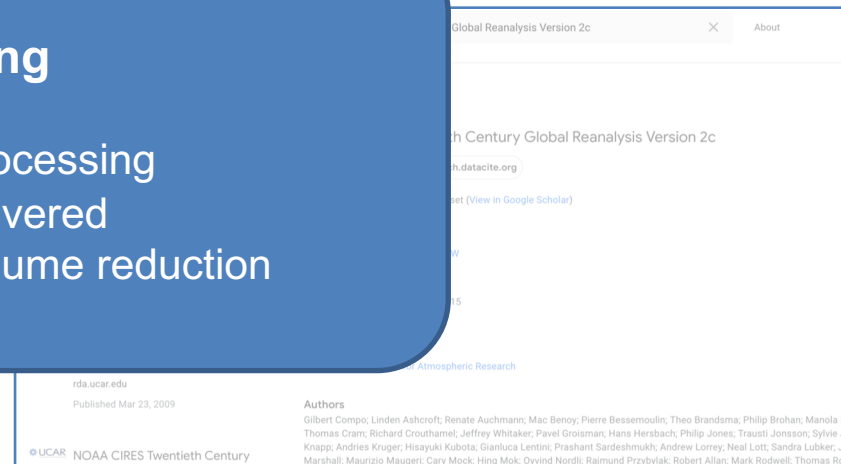
toolbox.google.com/datasetsearch

Metadata processing & support

- Maintain programmatic data ingest capabilities
- Support a range of search/discovery
 - Metadata Standards: DIF, FGDC, etc.
 - Controlled vocabularies
 - Access points
 - Schema.org metadata (to Google Search)
- Support data subsetting services
 - 97% data download reduction

FY 2018 data subsetting

- 57,000+ user requests
- 43 PB data input for processing
- 749 TB data output delivered
- 98% data download volume reduction




Lessons learned/Looking ahead



- Dataset file feature harvesting is critical
 - Supports value added services
 - Validates file integrity during data ingest
- Programmatic harvest of data citation metrics is just beginning.
 - Wanted:
 - More support from publishers through API access
 - Adherence to citation guidelines (authors)
 - Participate in community efforts to broaden this capability (e.g. Make Data Count, makedatacount.org)
- For more info: rda.ucar.edu/#!data-citation

Contact us

 [*rda.ucar.edu*](http://rda.ucar.edu)
 [*rdahelp@ucar.edu*](mailto:rdahelp@ucar.edu)



 [*@NCAR_RDA*](https://twitter.com/NCAR_RDA)

 [*@NCAR.RDA*](https://www.facebook.com/NCAR.RDA)

 [*ncarrda.blogspot.com*](http://ncarrda.blogspot.com)

Backup slides

DOI policies – RDA use cases

1. Create a new DOI dataset
2. Complete dataset replacement (new data from the provider)
3. Routine DOI dataset extension in time (regularly updated datasets)
4. Removal of DOI dataset
5. Small scale data replacement within a dataset (error fixes)