

Accessing Multi-TB-sized Datasets: Research Data Archive at NCAR

Douglas Schuster
Computational and Information Systems Laboratory
at NCAR

[\(schuster@ucar.edu\)](mailto:schuster@ucar.edu)

<http://rda.ucar.edu>





BUFF
No. 24908

Internet Access | Learning | Working | Teaching | Research | Outreach | 24 Hours

Highlights

- What is the RDA?
- Evolution of RDA Services
- User Identity Management
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Conclusions

Highlights

- What is the RDA?
- Evolution of RDA Services
- User Identity Management
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Conclusions

What is the RDA?

- NCAR/CISL Research Data Archive
- Purposes
 - Support climate & weather research at NCAR and UCAR Universities with reference datasets
- Collections
 - Ocean & atmosphere observations, analyses, reanalyses, operational NWP outputs
- Basic Metrics
 - Established in 1960s
 - 600+ datasets, 8M files, 1.8 PB
 - 70+ datasets growing daily – monthly

What is the RDA?

Science educated staff
 Expert consultants and data engineers
 Free and open access

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

<http://rda.ucar.edu>

[Home](#) | [Find Data](#) | [Ancillary Services](#) | [About/Contact](#) | [Data Citation](#) | [Web Services](#) | [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 |

Get Help:

- Frequently Asked Questions
- Reset your password
- A-Z Site Index
- RDA Users Email List
- Email Us

Recent News:

Guidance to WRF users for new NCEP GFS and FNL GRIB2 files
 rdahelp@ucar.edu has fielded many requests for assistance from WRF users attempting to use the new NCEP ...

New Dataset: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive
 The NCAR RDA now includes a historical archive of the operational NCEP GFS 0.25 degree ...

NCEP FNL dataset impacted by recent GFS overhaul
 Effective January 14, 2015, NCEP made some changes to their Global Forecast System (GFS). This impacts ds083.2 ...

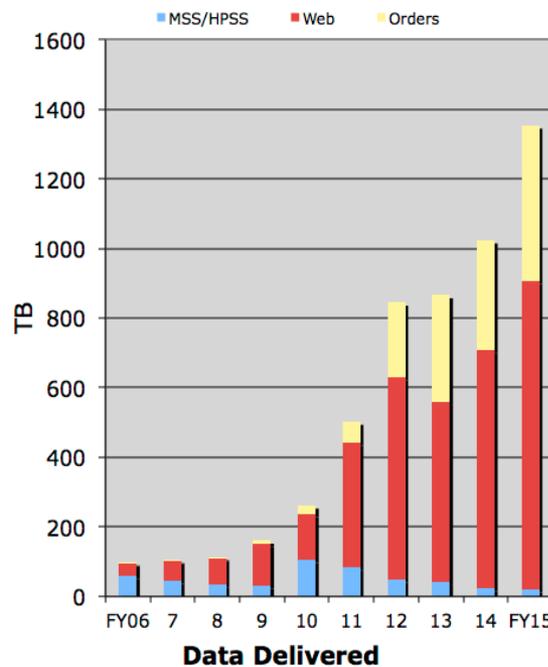
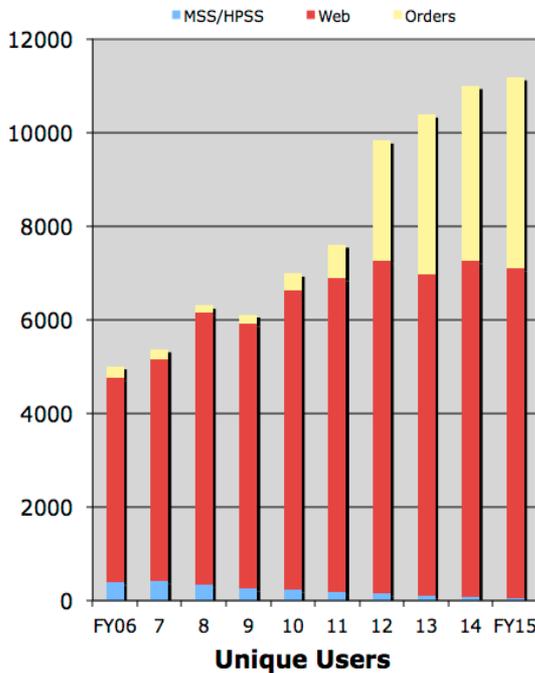
[Archive...](#)

Recently Added Datasets: (within the last 6 months)

- NCEP GFS 0.25 Degree Global Forecast Auxiliary "b" Grids Historical Archive
- Cloud Properties from ISCCP and PATMOS-x Corrected for Spurious Variability Related to Changes in Satellite Orbits, Instrument Calibrations, and Other Factors
- NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive
- NOAA CPC Morphing Technique (CMORPH) Global Precipitation Analyses Version 0.x (June 2014 - current)
- NCAR CESM Global Bias-corrected CMIP5 Output to Support WRF/MPAS Research
- ERA-20C Project (ECMWF Atmospheric Reanalysis of the 20th Century)
- WCRP and WWRP THORPEX YOTC (Year of Tropical Convection) Project, Single Parameter 3-Hourly Pressure Level Forecast Time Series, Transformed to a Regular 1600 by 800 (N400) Gaussian Grid, Dynamical Parameters Only

Other Ways to Explore:

- GCMD Topic:
 - Agriculture • Atmosphere • Biosphere • Climate Indicators • Cryosphere • Hydrosphere • Land Surface • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • Sun-earth Interactions



Highlights

- What is the RDA?
- **Evolution of RDA Services**
- User Identity Management
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Conclusions

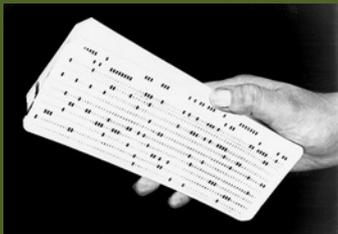
Evolution of RDA services

- Systematically maintain a large diverse community archive
- Make data discovery “easy”
- Improve access to serve many more users
- Support reproducible research
- Reduce the time researchers spend dealing with data

Evolution of RDA services

- 1960s to early 1990s – **Data Consultant Driven**
 - One to one service model

Data/Metadata Storage



Data Discovery



Data Delivery



Evolution of RDA services

- Mid 1990s to Early 2000s – Data Consultant Driven

Data/Metadata Storage



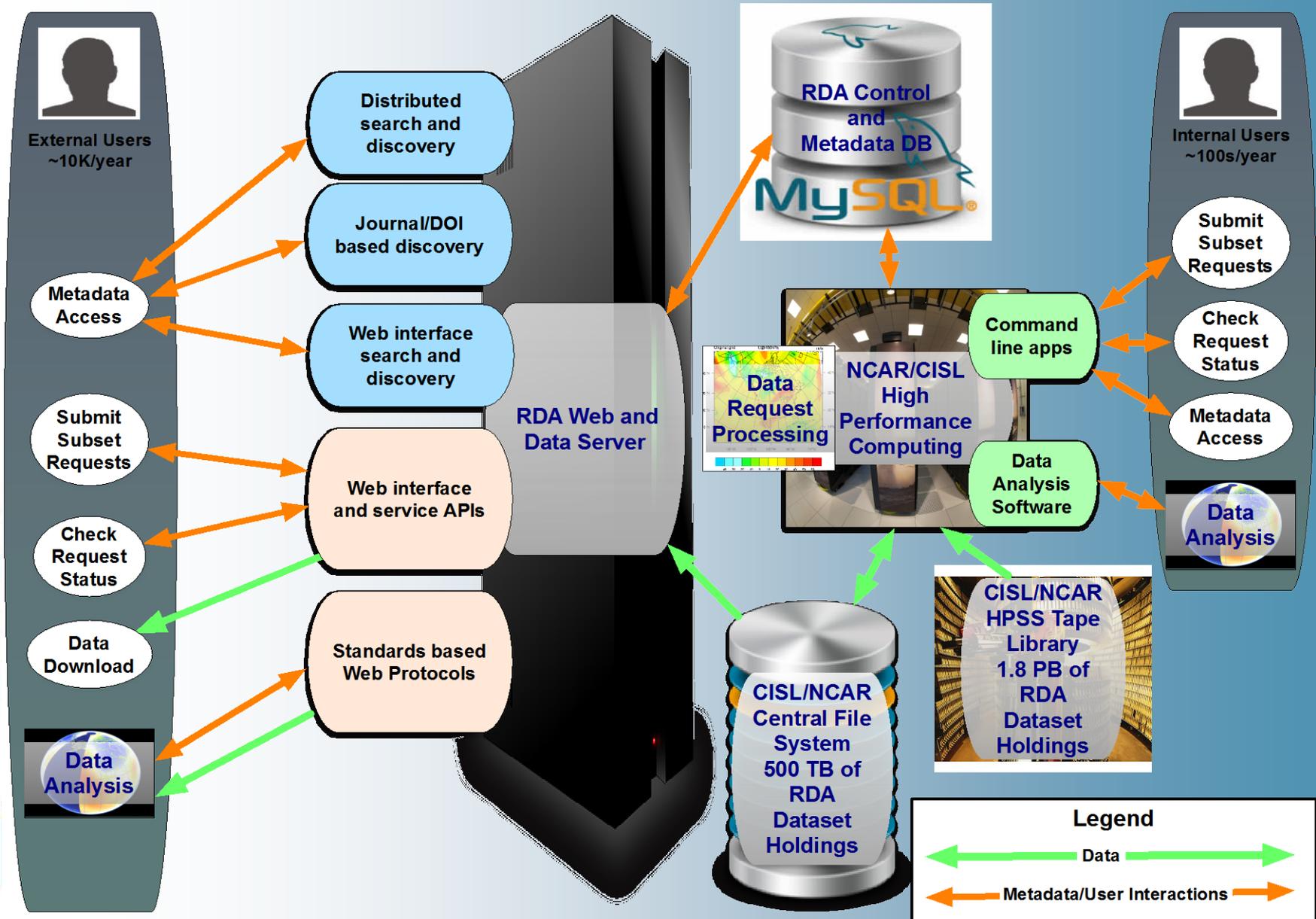
Data Discovery



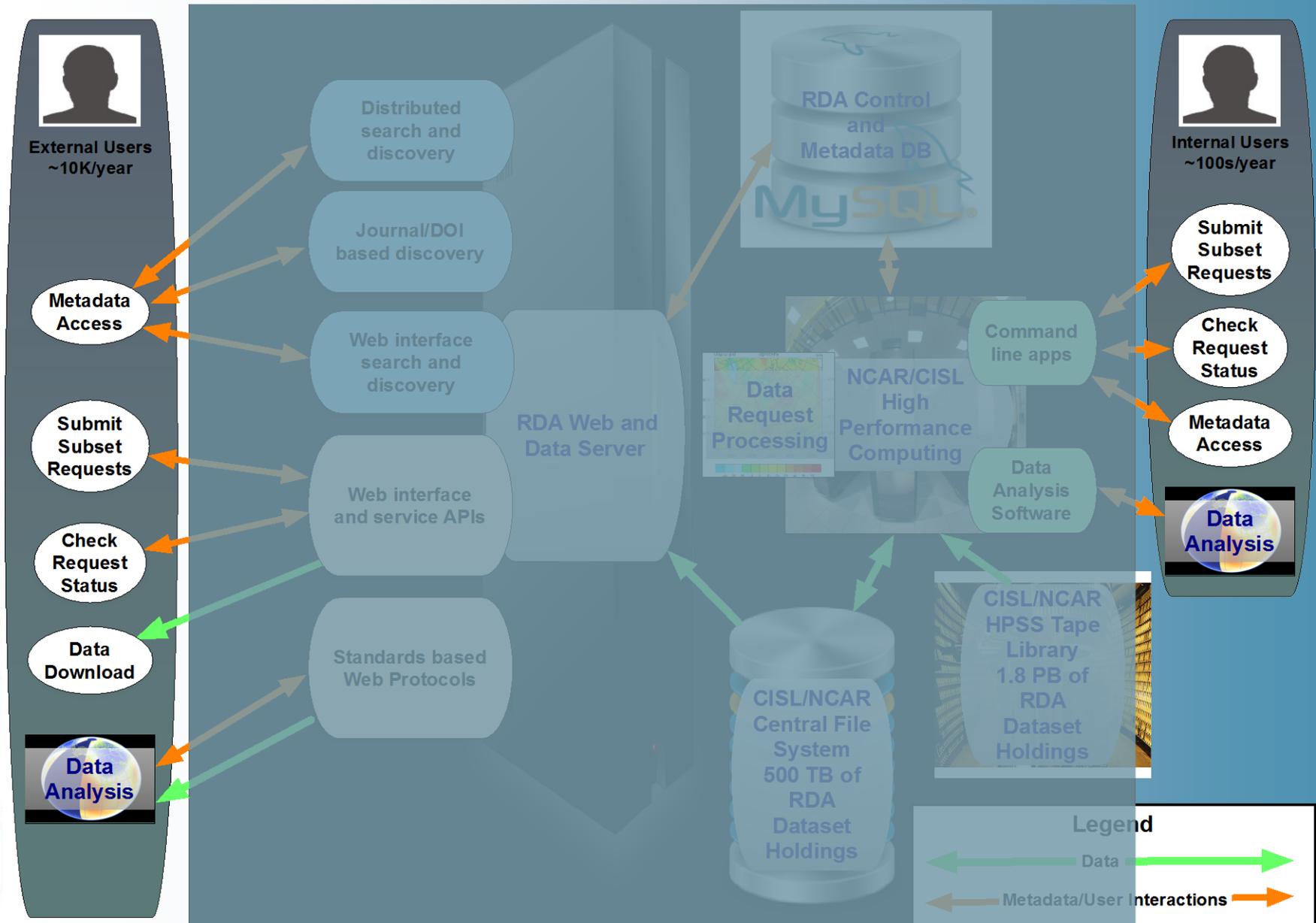
Data Delivery



Current RDA Services



Current RDA Services



RDA Data Discovery Pathways



RDA Data Discovery Pathways



RDA Data Discovery Pathways

-Server Generated Data Citation

How to Cite This Dataset:

RIS

BibTeX

Compo, G. P., and Coauthors, 2009: NOAA CIRES Twentieth Century Global Reanalysis Version 2. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. [Available online at <http://dx.doi.org/10.5065/D6QR4V37>.] 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](#)

Dataset:

NOAA CIRES Twentieth Century Global Reanalysis Version 2² (ds131.1)

Your Access History for July 2013:

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.

| July 2013 | | | | | | |
|-----------|-----|-----|----------|-----|-----|-----|
| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
| | 1 | 2 | <u>3</u> | 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 | | | |

Maintain user access history, citation recall at any time

For Data Accessed on 2013-07-03:

Dataset Citation: **RIS**

Compo, G. P., et al. 2009. *NOAA CIRES Twentieth Century Global Reanalysis Version 2*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. <http://dx.doi.org/10.5065/D6QR4V37>. Accessed 3 Jul 2013.

Bibliographic citation shown in style

Data Access Detail:

1 subset request:

- 18 files, 3.44 MB

Date Limits : 1975-06-15 00:00 to 1975-06-30 12:00

Parameter : TMP

Level Type : UGR2

RDA Data Discovery Pathways



RDA Data Discovery Pathways

-Advise User of Best Dataset Collection(s) for purpose

NCAR UCAR ClimateDataGuide *inform • compare • discover*

CLIMATE DATA ANALYSIS TOOLS MODEL EVALUATION EXPERT CONTRIBUTORS ABOUT Site-wide Search >>

Data Discovery Guided by Experts >>

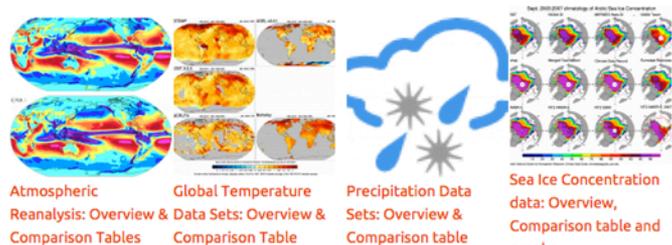
Search and access 174 data sets covering the Atmosphere, Ocean, Land and more. Explore climate indices, reanalyses and satellite data and understand their application to climate model metrics. This is the only data portal that combines data discovery, metadata, figures and world-class expertise on the strengths, limitations and applications of climate data. [Discover it now.](#)

See data pages with guidance from these experts:



Data Set Overviews >>

Compare the attributes, strengths and limitations of multiple data sets.



Atmospheric Reanalysis: Overview & Comparison Tables
Global Temperature Comparison Table

Precipitation Data Sets: Overview & Comparison table

Sea Ice Concentration data: Overview, Comparison table and graphs

JOIN US

Multiply the impact of your work by announcing new data sets and sharing your knowledge of the strengths, limitations and applications of particular data sets.

Ways to make an impact

- Become a registered user of this site
- Contribute a data set & assessment
- Post a comment to any data set page

SHARE CLIMATE DATA GUIDE

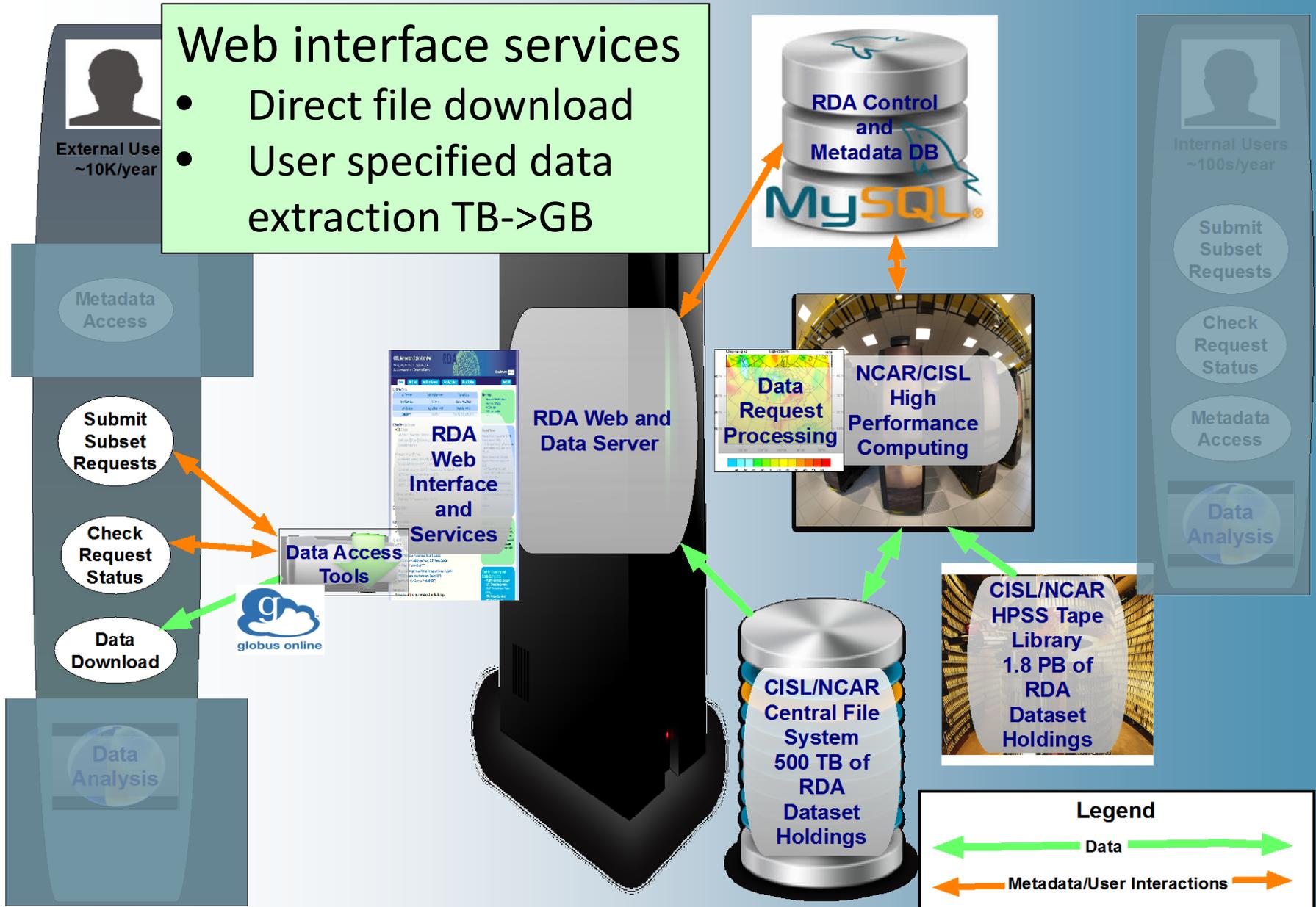


FOLLOW US FOR UPDATES

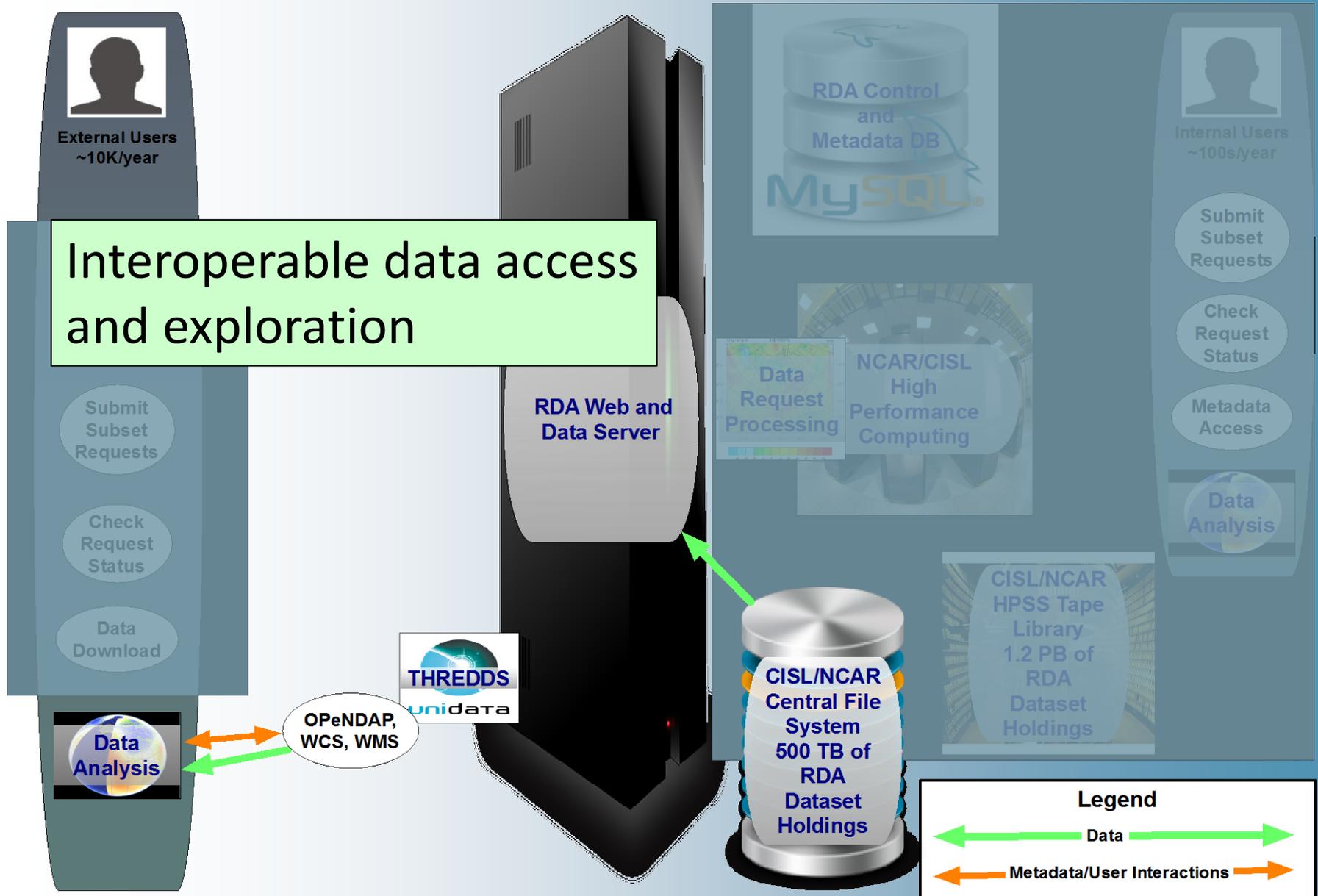
Follow us on Twitter for content updates & relevant data news.




RDA Data Access Pathways



RDA Data Access Pathways



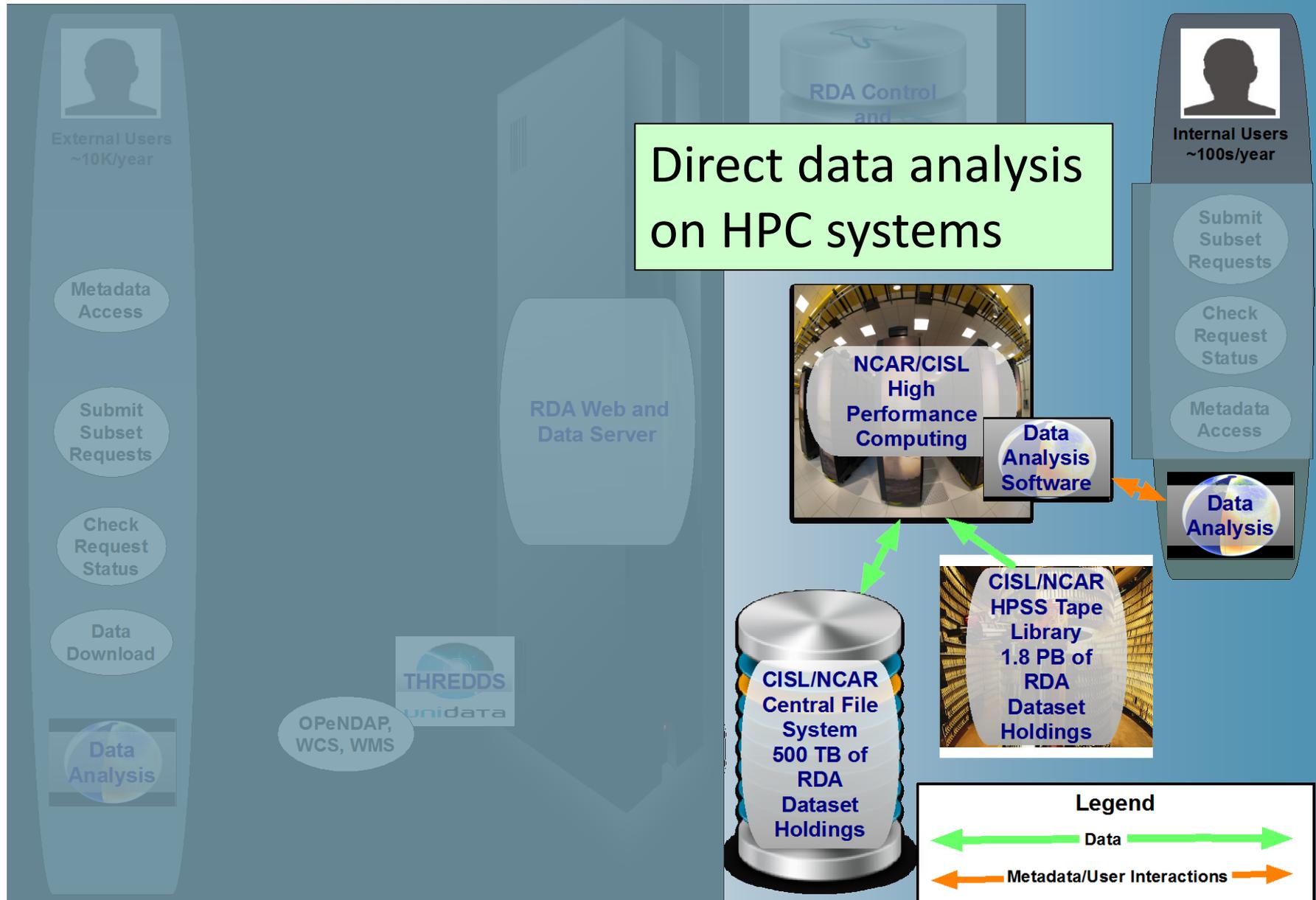


ratory

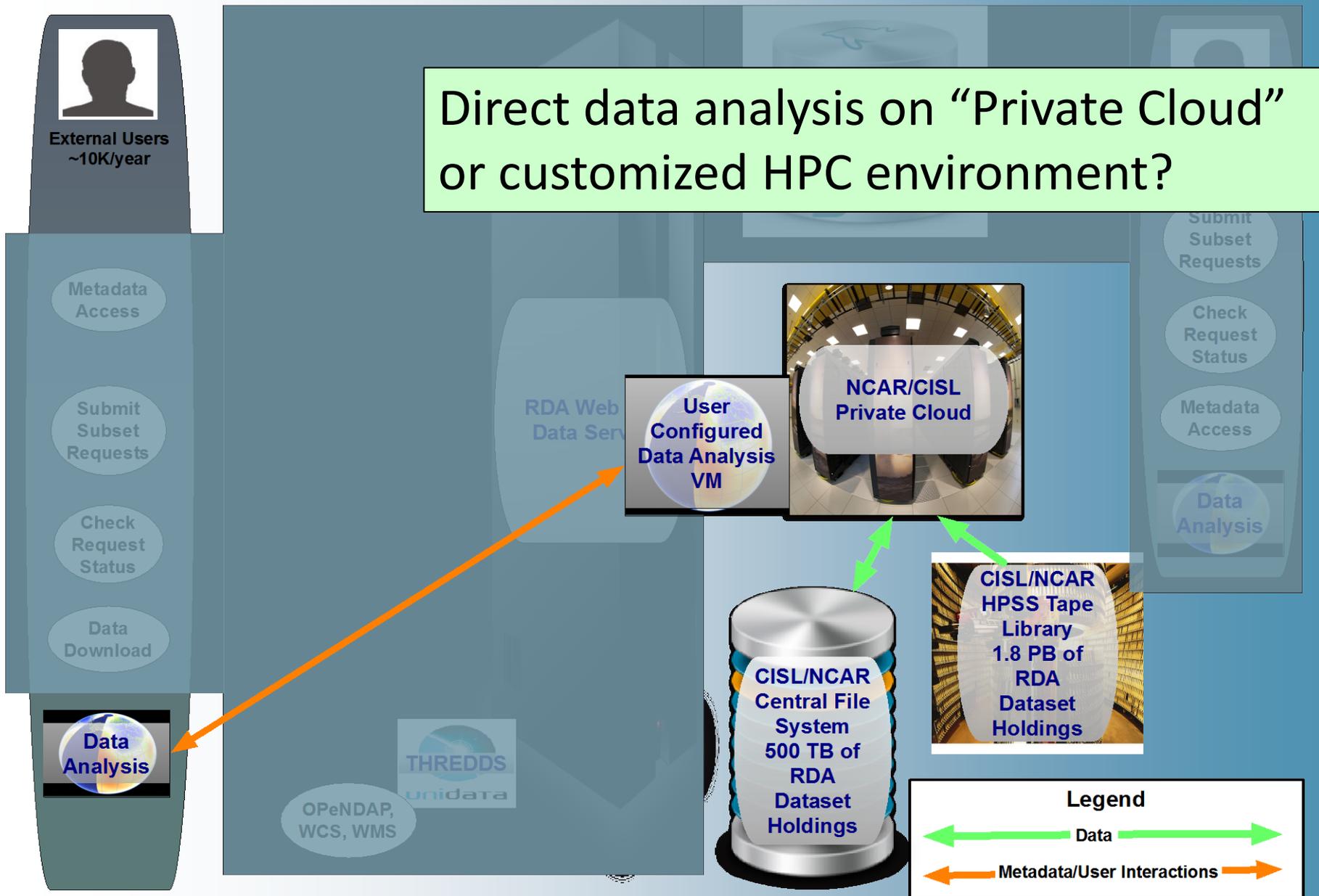
OPeNDAP Access Example Product



RDA Data Access Pathways



Future RDA Data Access Pathways



Highlights

- What is the RDA?
- Evolution of RDA Services
- **User Identity Management**
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Conclusions



User Identity Management

- Currently many identity management solutions across NCAR
 - RDA (home grown)
- Need a “single-sign-on” and alt-ID solution
 - Ease of use
 - Improve security
 - Reduced costs –eliminate duplication of effort
 - Integrated app development
- User identity as a service?
 - Globus Auth and Nexus?





User Identity Management

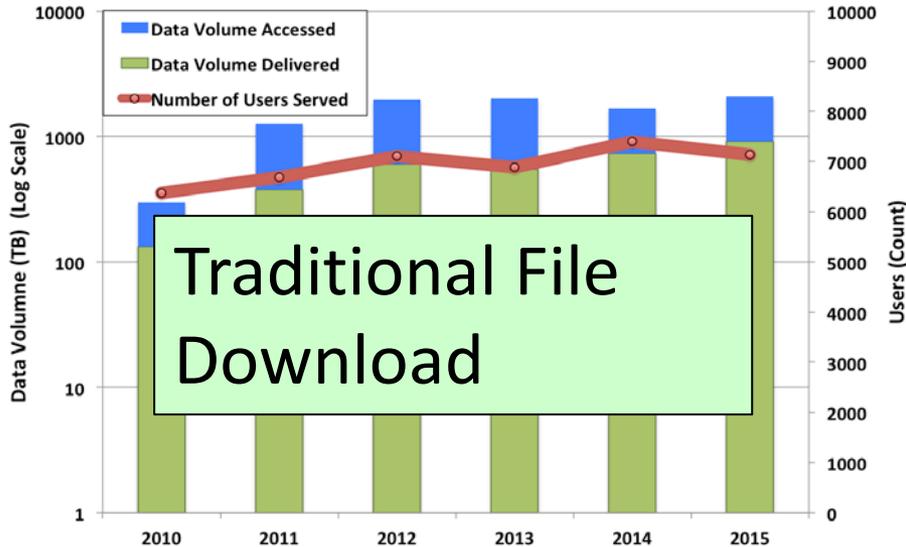
- Globus Auth
 - Remove the Need for an RDA Identity
 - Support login with Globus Auth federated identity providers
 - University login
 - Social accounts (google, facebook)
 - Globus Auth sends a standard OpenID connect token that identifies the user
- Globus Nexus
 - Manage RDA Identities?

Highlights

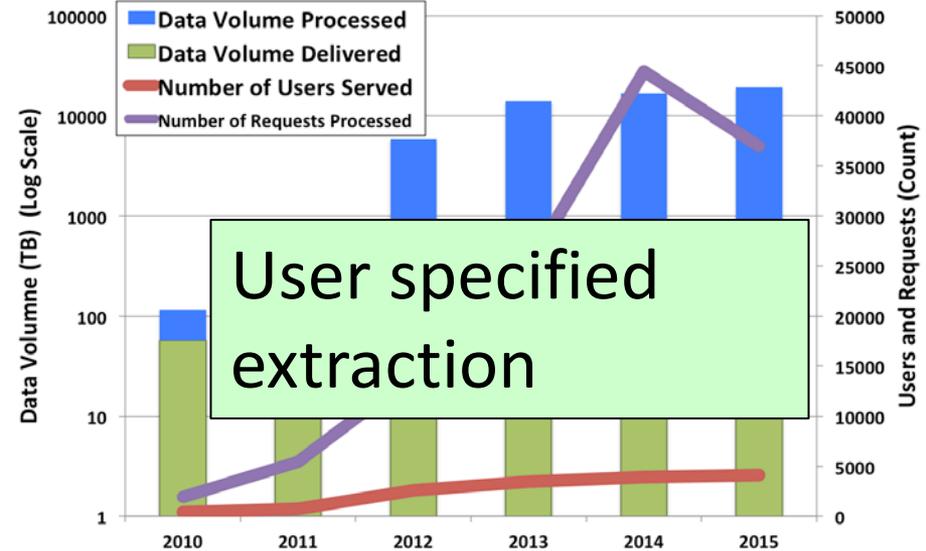
- What is the RDA?
- Evolution of RDA Services
- User Identity Management
- **Usage Metrics**
- User Outreach and User Support
- Lessons Learned and Conclusions

Current Services Overview – Usage metrics

Yearly RDA User Access from Web Interface
Direct Archive File Downloads

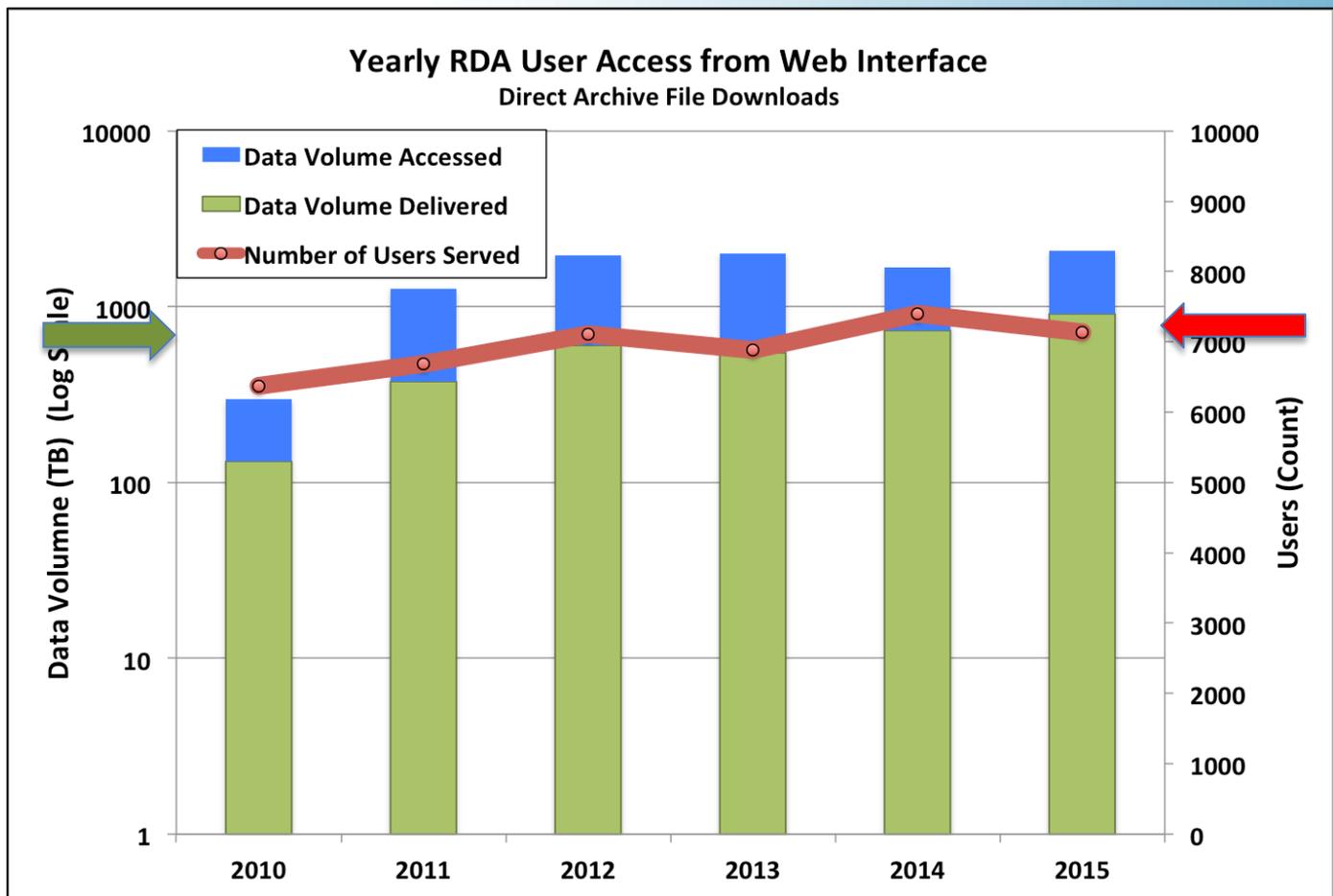


Yearly Customized RDA User Access from Web Interface
Automated Archive Subsetting and Format Conversion Requests





Traditional File Download – Usage metrics



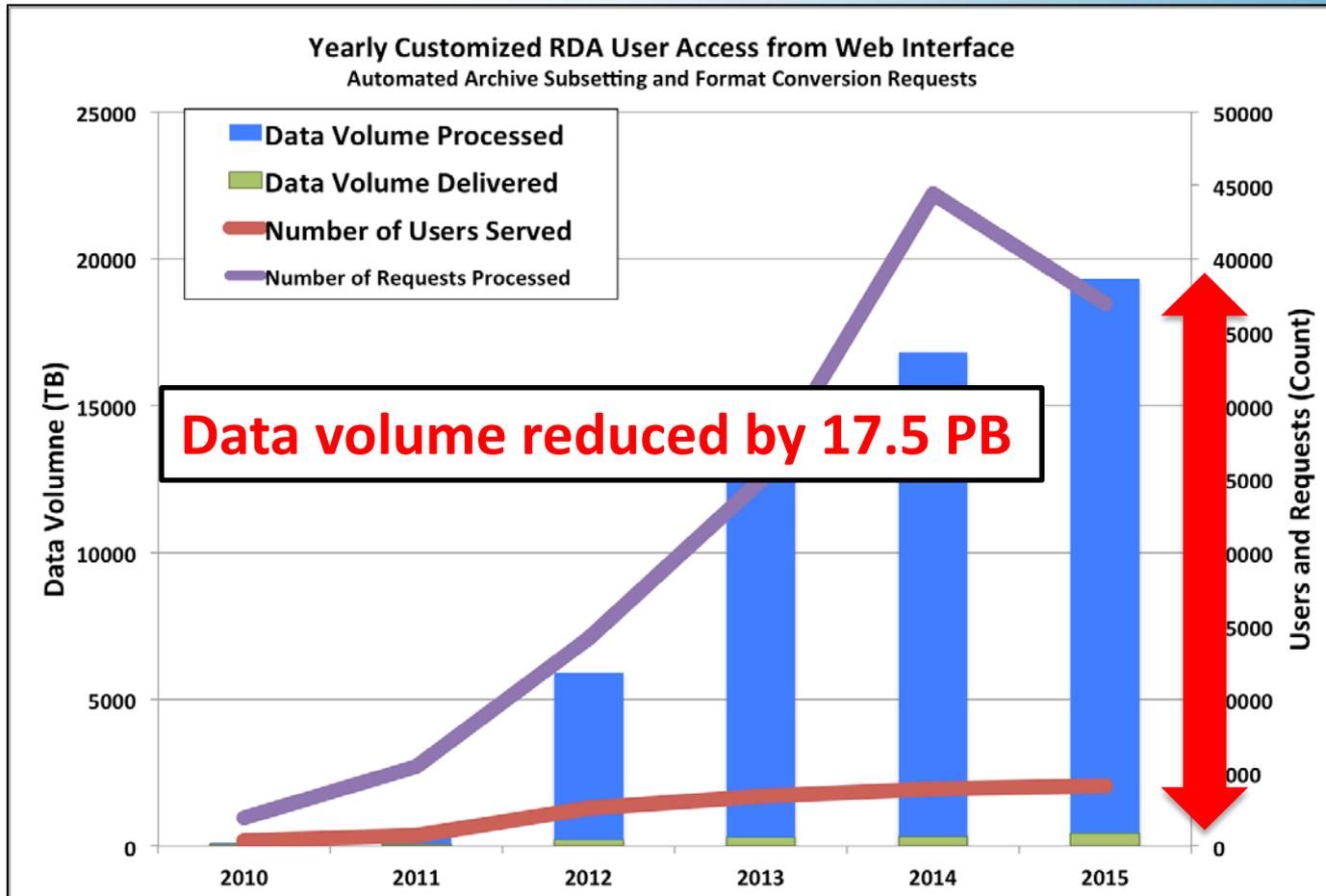
In FY 2015 (through Sep 1)

- 7100 users
- 900TB of data delivered





User Specified Extraction –Usage Metrics

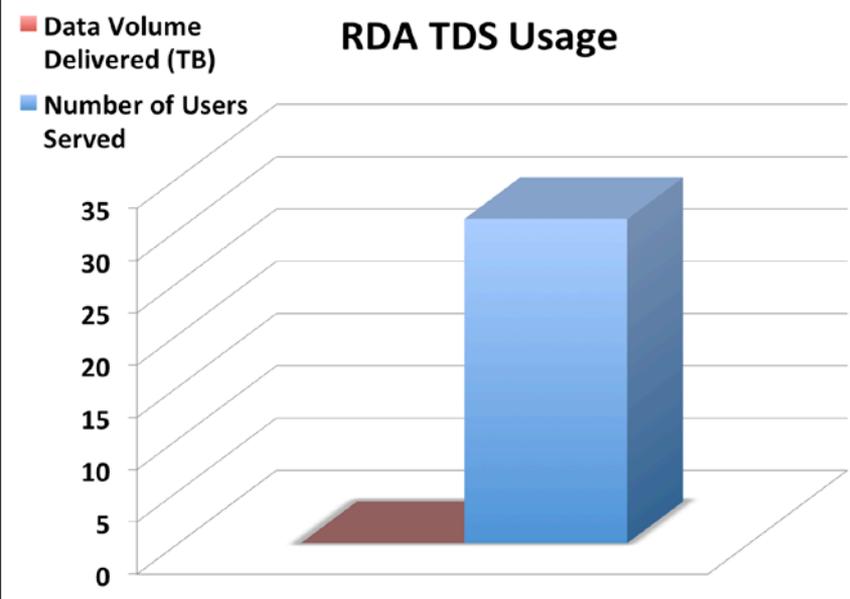
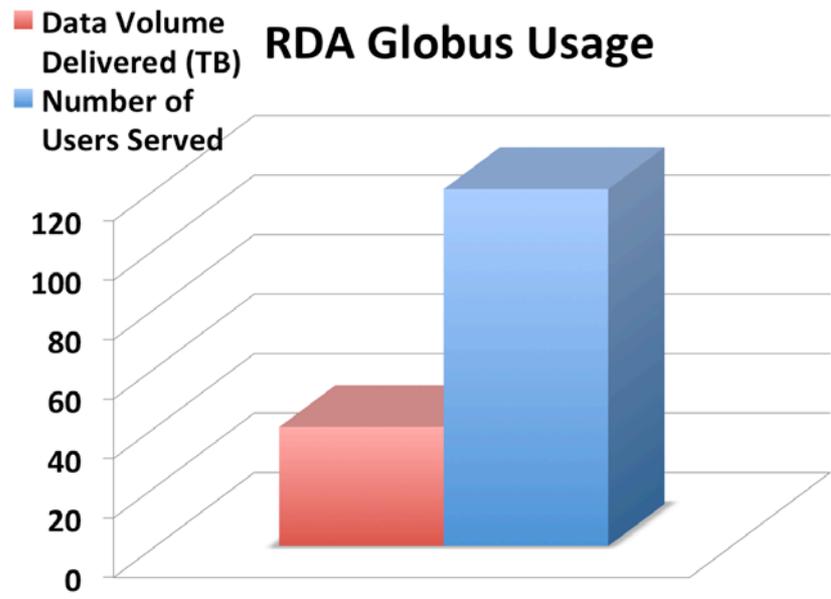


In FY 2015 (through Sep 1)

- 37K user requests
- 4000+ users submitted data extraction requests
- 18+ PB data processed (**“Parallel” processing on HPC**)
- 380 TB of data delivered



New Services Overview –Usage metrics



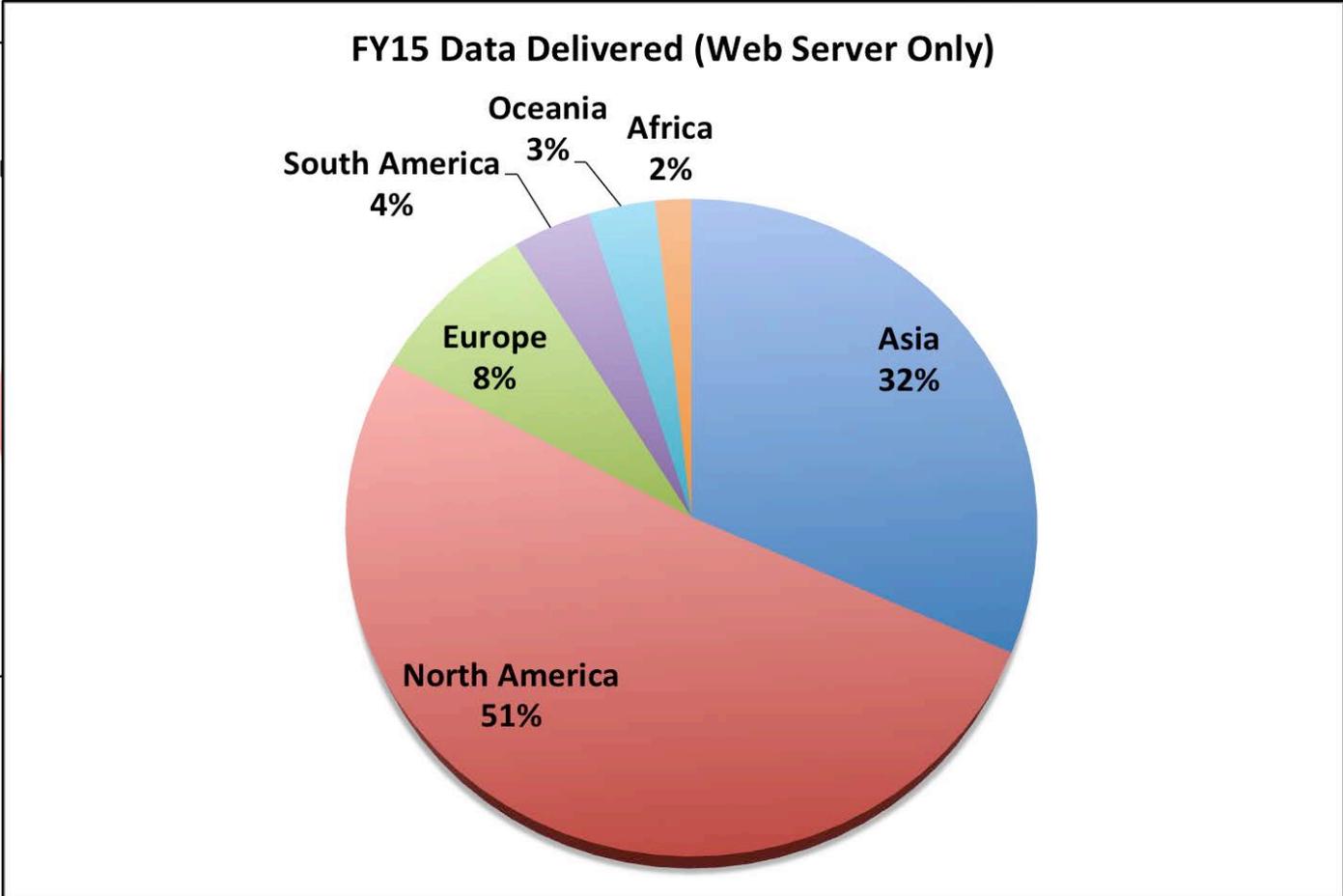
In FY 2015 (TDS and Globus usage)

- 120 users transferred 40 TB of data using Globus
- 31 users downloaded 1 TB of data using TDS services (OPeNDAP, Aug 2015)



Metrics: Users and Data Delivery by Region

Excluding HPC Usage at NCAR, Web services only



In FY 2015

- 51 % of unique users originate from Asia
- 51 % of data volume downloaded to North American users

Highlights

- What is the RDA?
- Evolution of RDA Services
- User Identity Management
- Usage Metrics
- **User Outreach and User Support**
- Lessons Learned and Conclusions

User Outreach and Support

is Laboratory

NCAR Research Data Archive Blog

News and tutorials from the National Center for Atmospheric Research's Research Data Archive.

Blog Archive

2015 (25)

August (5)

Interoperable Access to NCAR Research Data Archive...

Accessing RDA collections from Yellowstone

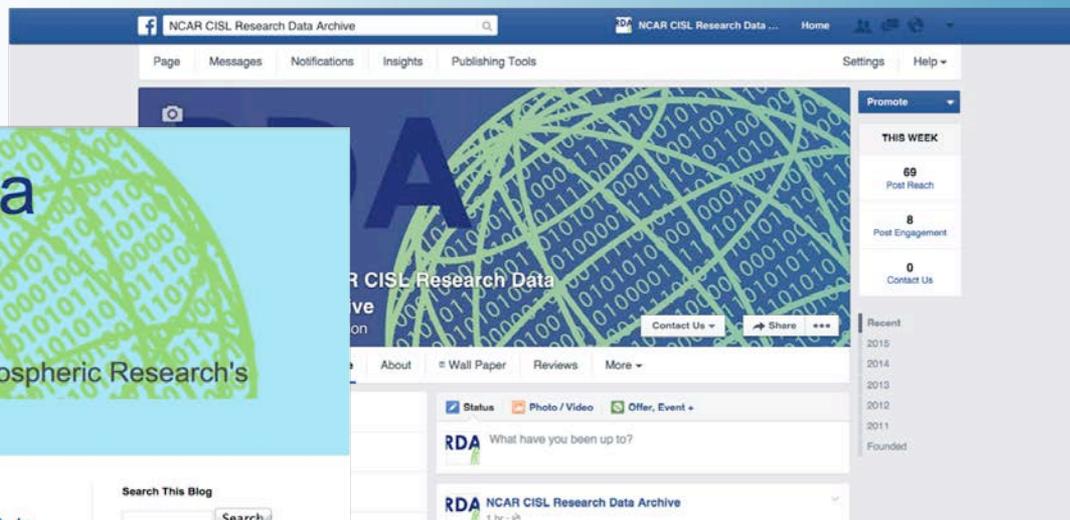
Cite your data

New Terms and Conditions of Use

2015-08-18

Interoperable Access to NCAR Research Data Archive Collections

To enhance user experience and utility, the RDA now offers THREDDS Data Server (TDS) access for many highly valued dataset collections (<http://rda.ucar.edu/thredds>). TDS offers datasets with native formats in GRIB1 and GRIB2 are present aggregations, enabling users to access an entire dataset (that can be comprised of 1000's of files, through a single v



Comput



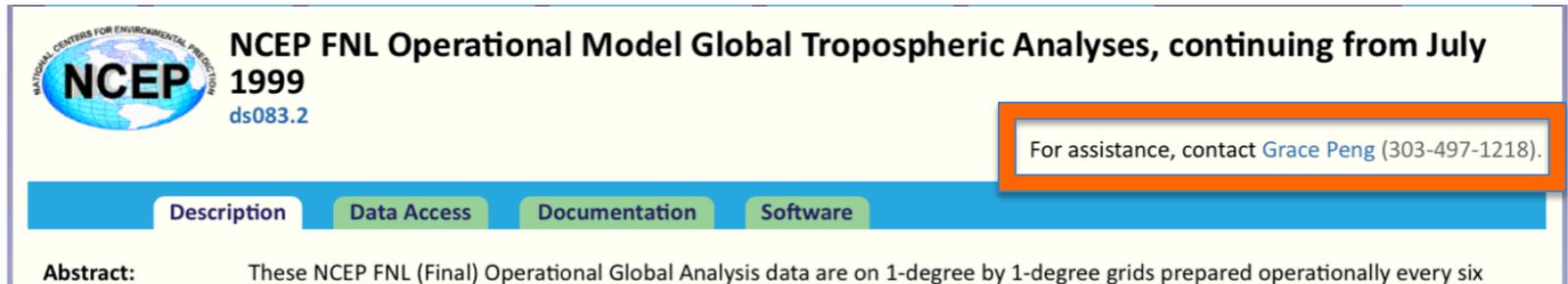
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH





User Outreach and Support

- Dataset consultant listed on each dataset homepage



The screenshot shows the homepage for the NCEP FNL Operational Model Global Tropospheric Analyses dataset. The header includes the NCEP logo and the text "NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999 ds083.2". A navigation bar contains links for "Description", "Data Access", "Documentation", and "Software". An orange-bordered box highlights the contact information: "For assistance, contact Grace Peng (303-497-1218)". The abstract text reads: "These NCEP FNL (Final) Operational Global Analysis data are on 1-degree by 1-degree grids prepared operationally every six

- General help email rdahelp@ucar.edu
- Future: User tutorial video series

Highlights

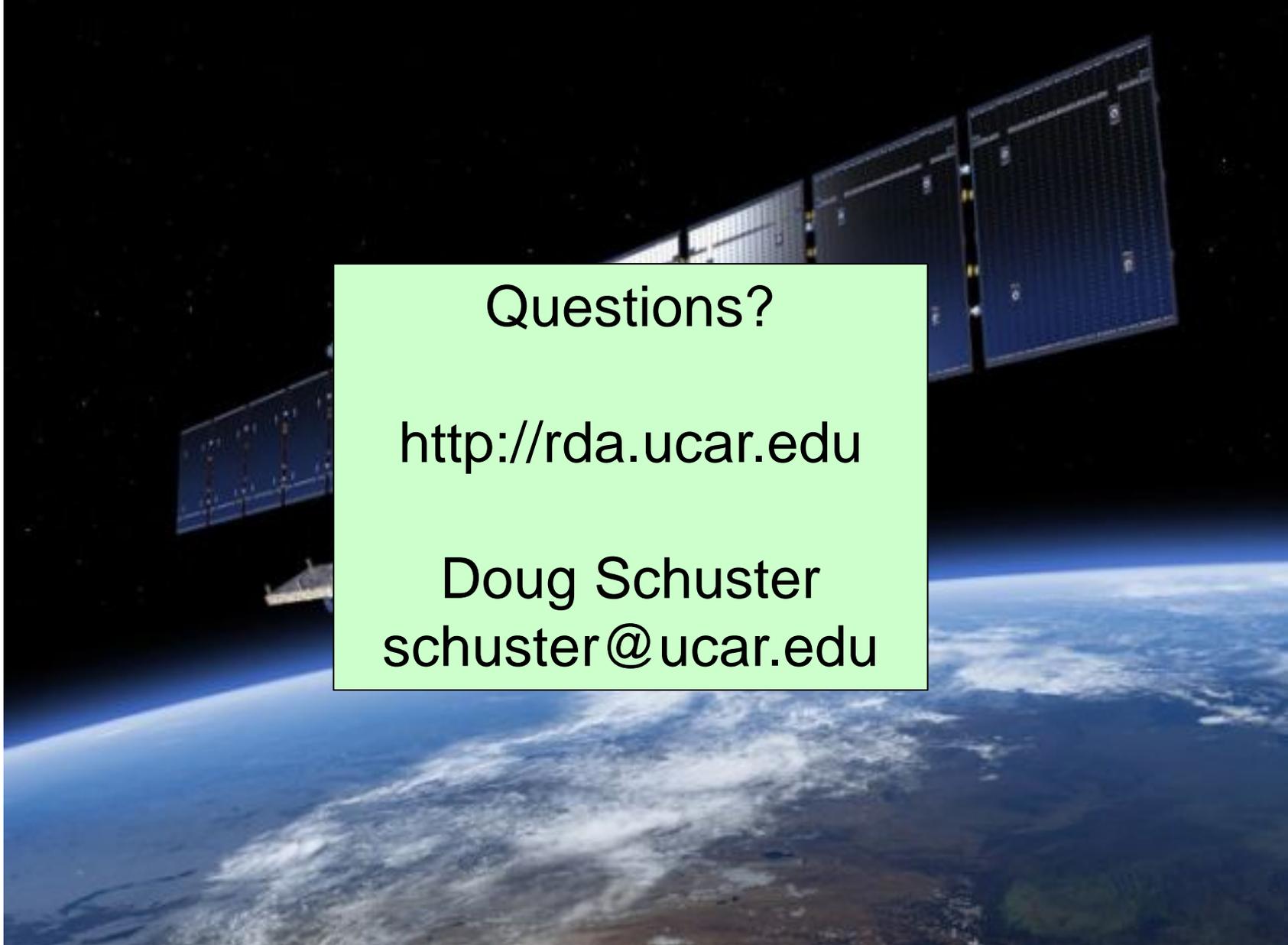
- What is the RDA?
- Evolution of RDA Services
- User Identity Management
- Usage Metrics
- User Outreach and User Support
- **Lessons Learned and Conclusions**

Lessons Learned

- Archive structure impacts efficiency of data access services
 - Use cases: climate vs weather research
 - Use of tar packages/compression
- Programmatic metadata harvesting critical to support value added services
- Faceted search tends to be overly complicated and not regularly used
- Interoperable tools lack authentication support
- Usage metrics for HPC users tough to capture

Conclusions

- The RDA is a large, growing, heterogeneous archive
- Evolution of RDA Services
 - Consultant Driven -> One-to-one
 - Metadata Driven -> One-to-many
- User Identity Management
 - Need single-sign-on and alt-id provider support
- User Outreach and User Support
 - Traditional methods and social media



Questions?

<http://rda.ucar.edu>

Doug Schuster
schuster@ucar.edu