# Accessing Multi-TB-sized Datasets at NCAR's Research Data Archive

**Douglas Schuster**, **Steven Worley**
Database Engineer, CISL/DSS
http://rda.ucar.edu

**ECMWF 16th MOS Workshop, Reading, UK**
March 1, 2017

# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# Overview of NCAR

## National Center for Atmospheric Research -1960

**Mission Statement:**

- To understand the behavior of the atmosphere and related Earth and geospace systems
- To support, enhance, and extend the capabilities of the university community and the broader scientific community, nationally and internationally, and
- To foster the transfer of knowledge and technology for the betterment of life on Earth

# Overview of NCAR

- Federally Funded Research and Development Center
  - National Science Foundation
  - Other government agencies, other national governments, private sector
- Managed by University Corporation for Atmospheric Research (UCAR)
- 5 campuses, 1200+ employees.

# Overview of the RDA

Research Data Archive

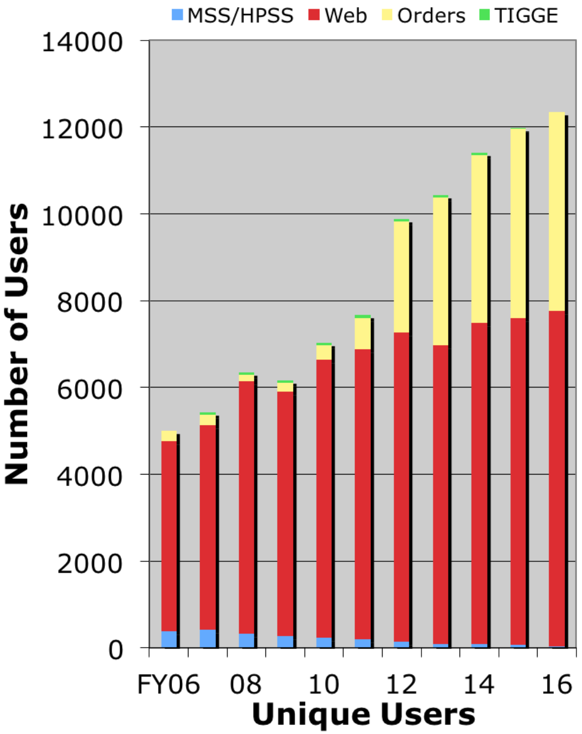
- **NCAR’s Research Data Archive**
- **Purposes**
  - Support climate & weather research at NCAR and UCAR universities with reference datasets
- **Collections**
  - Ocean & atmospheric observations, analyses, reanalyses, operational NWP products
  - Established in 1960s
  - 700+ datasets, 8M files, 1.8 PB
  - 70+ datasets growing daily-monthly
- **Free and open access**
- **Worldwide usage**
- **Science educated staff**
- **One of many data assets at NCAR**

# Overview of the RDA

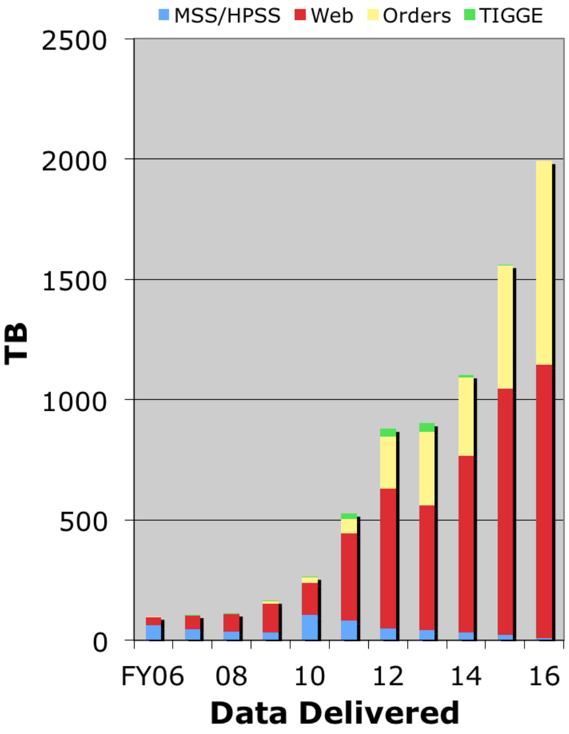
Research Data Archive

12,000+ Unique Web users in FY 2016

~2 PB Data Delivered in FY 2016

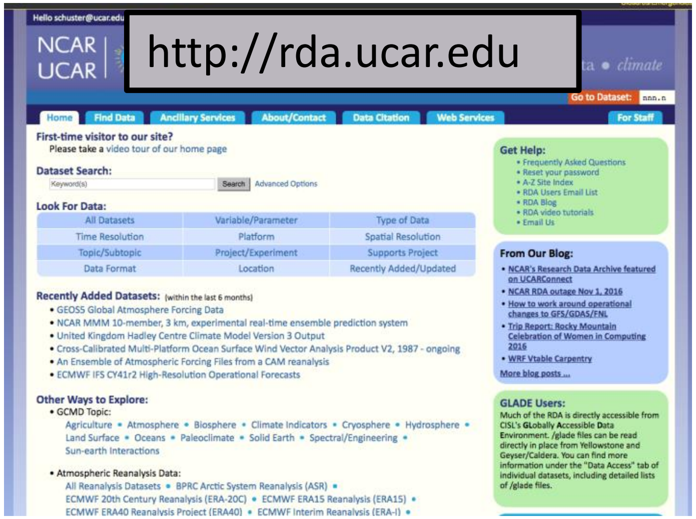


Unique Users



Data Delivered

http://rda.ucar.edu



**First-time visitor to our site?**
Please take a video tour of our home page

**Dataset Search:**

**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 |

**Recently Added Datasets:** (within the last 6 months)

- GEOS5 Global Atmosphere Forcing Data
- NCAR MMM 10-member, 3 km, experimental real-time ensemble prediction system
- United Kingdom Hadley Centre Climate Model Version 3 Output
- Cross-Calibrated Multi-Platform Ocean Surface Wind Vector Analysis Product V2, 1987 - ongoing
- An Ensemble of Atmospheric Forcing Files from a CAM reanalysis
- ECMWF IFS CY41r2 High-Resolution Operational Forecasts

**Other Ways to Explore:**

- GCMD Topic:
  Agriculture • Atmosphere • Biosphere • Climate Indicators • Cryosphere • Hydrosphere • Land Surface • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • Sun-earth Interactions
- Atmospheric Reanalysis Data:
  All Reanalysis Datasets • BPRC Arctic System Reanalysis (ASR) • ECMWF 20th Century Reanalysis (ERA-20C) • ECMWF ERA15 Reanalysis (ERA15) • ECMWF ERA40 Reanalysis Project (ERA40) • ECMWF Interim Reanalysis (ERA-I) •

**Get Help:**

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

**From Our Blog:**

- NCAR's Research Data Archive featured on UCARConnect
- NCAR RDA outage Nov 1, 2016
- How to work around operational changes to GFS/GDAS/FNL
- Trip Report: Rocky Mountain Celebration of Women in Computing 2016
- WRF Vtable Carpentry

More blog posts ...

**GLADE Users:**
Much of the RDA is directly accessible from CISL's **GL**obally **A**ccessible **D**ata **E**nvironment. /glade files can be read directly in place from Yellowstone and Geyser/Caldera. You can find more information under the "Data Access" tab of individual datasets, including detailed lists of /glade files.

# Overview of the RDA Collaboration Examples

ECMWF

気象庁 Japan Meteorological Agency

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION U.S. DEPARTMENT OF COMMERCE

NASA

中国气象局 CHINA METEOROLOGICAL ADMINISTRATION

- Host Datasets and Share Data
  - Operational model output
  - Reanalysis products
  - Observations to support reanalysis efforts
- Develop Products and Services
  - E.G. ICOADS ,ISPD, and NCAR UA soundings
  - Interpolate and restructure products

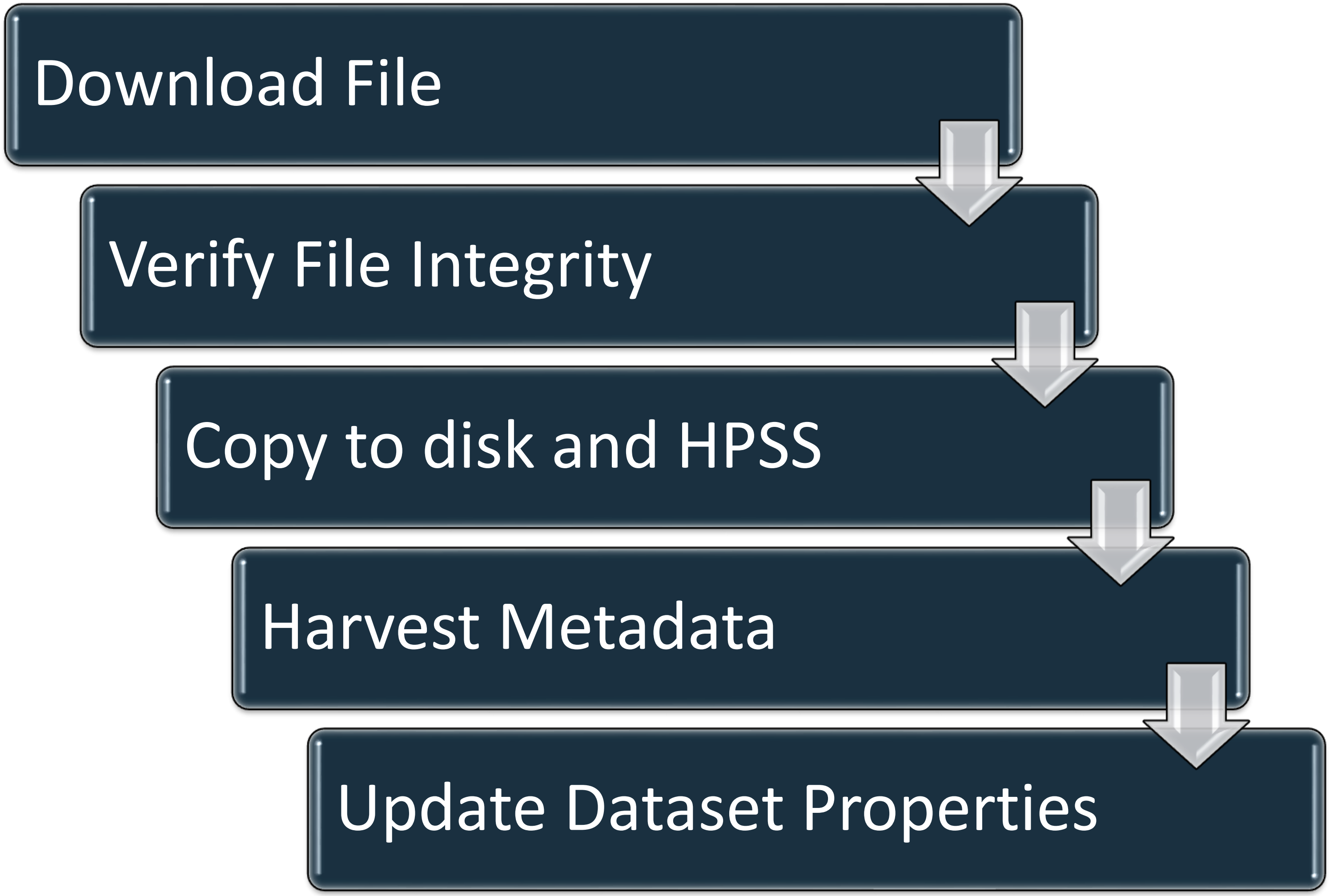
# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# Data Management Strategies

- Develop and leverage tools that support programmatic data ingest
  - Support scalable archive growth and maintenance
- Establish and maintain metadata databases
  - Support discovery, access, and value added services
- Store data at multiple physical locations
  - Support data security

# Data Management Strategies
# Typical Dataset Update Workflow



Download File

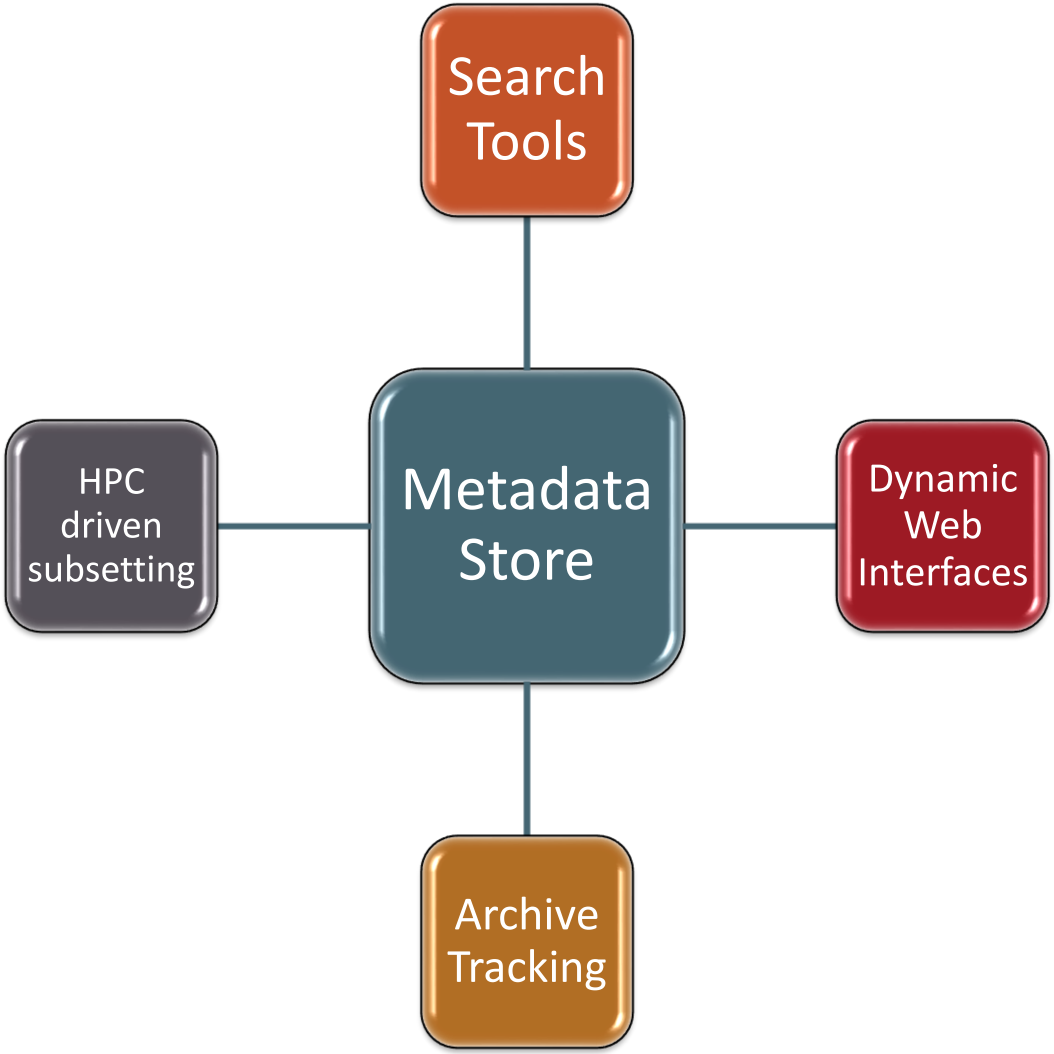
Verify File Integrity

Copy to disk and HPSS

Harvest Metadata

Update Dataset Properties

# Data Management Strategies
# Metadata is foundational



# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA Users Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# Metadata Overview

- **Metadata**
  - Dataset summary metadata stored in native RDA schema with GCMD control vocabulary
  - Exported to
    - ISO 19139, DIF, FGDC, OAI_DC, DATACITE, THREDDS
  - Shared via
    - OAI-PMH endpoint
    - CS/W endpoint
  - Harvested by
    - NCAR data search and discovery system
    - NASA Global Change Master Directory (GCMD)
    - Thompson Reuters
    - GEOSS –GEO portal

# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

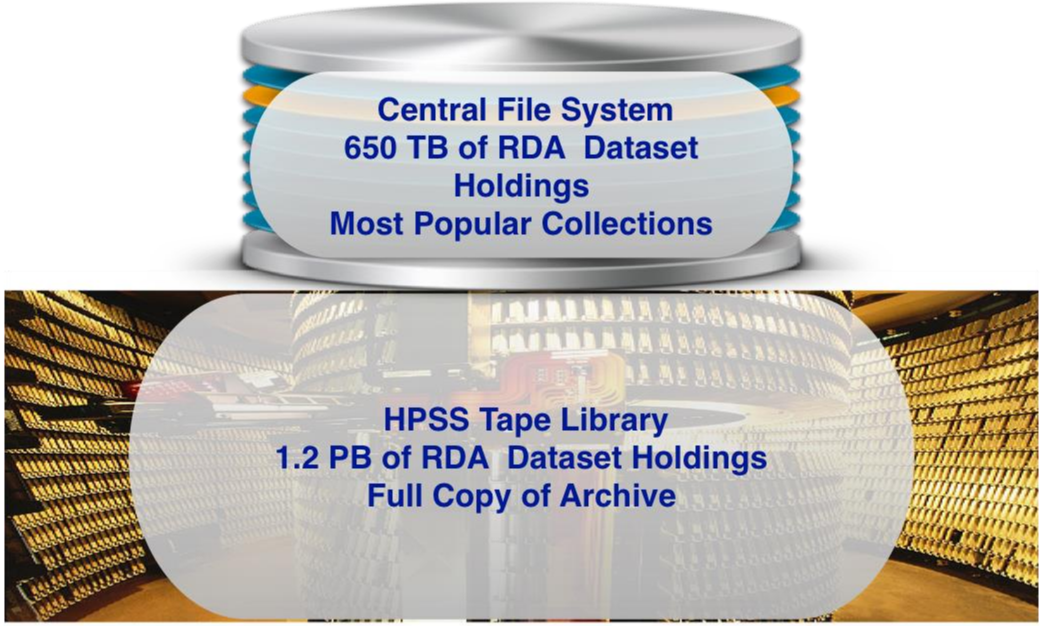
# RDA User Services

- Make data discovery “easy”
- Improve access to serve growth in user numbers and types
- Support reproducible research
  - DOI’s on 79 static and dynamic datasets
- **Reduce the time researchers spend dealing with data**

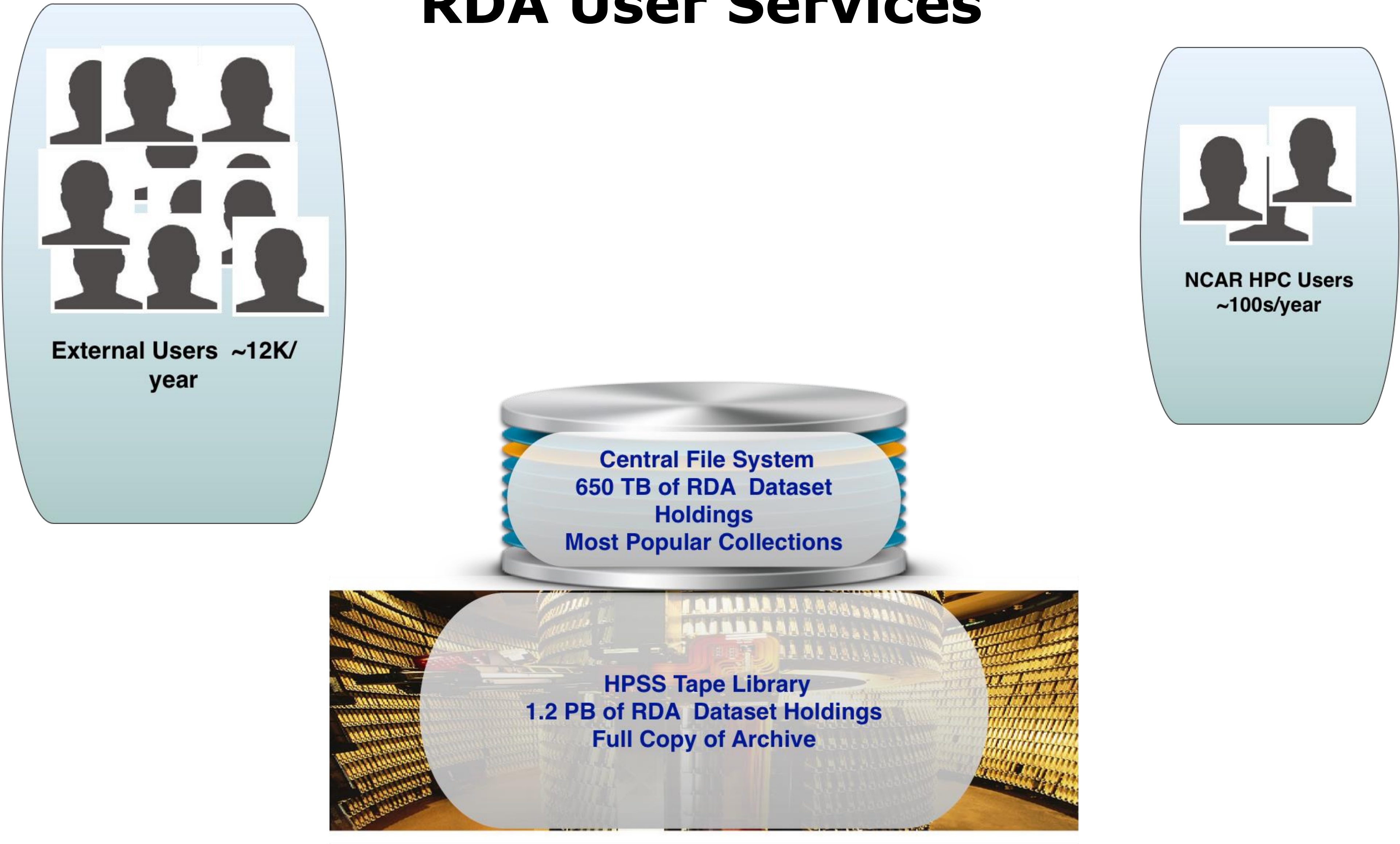
# RDA User Services

HPSS Tape Library
1.2 PB of RDA  Dataset Holdings
Full Copy of Archive

# RDA User Services



# RDA User Services



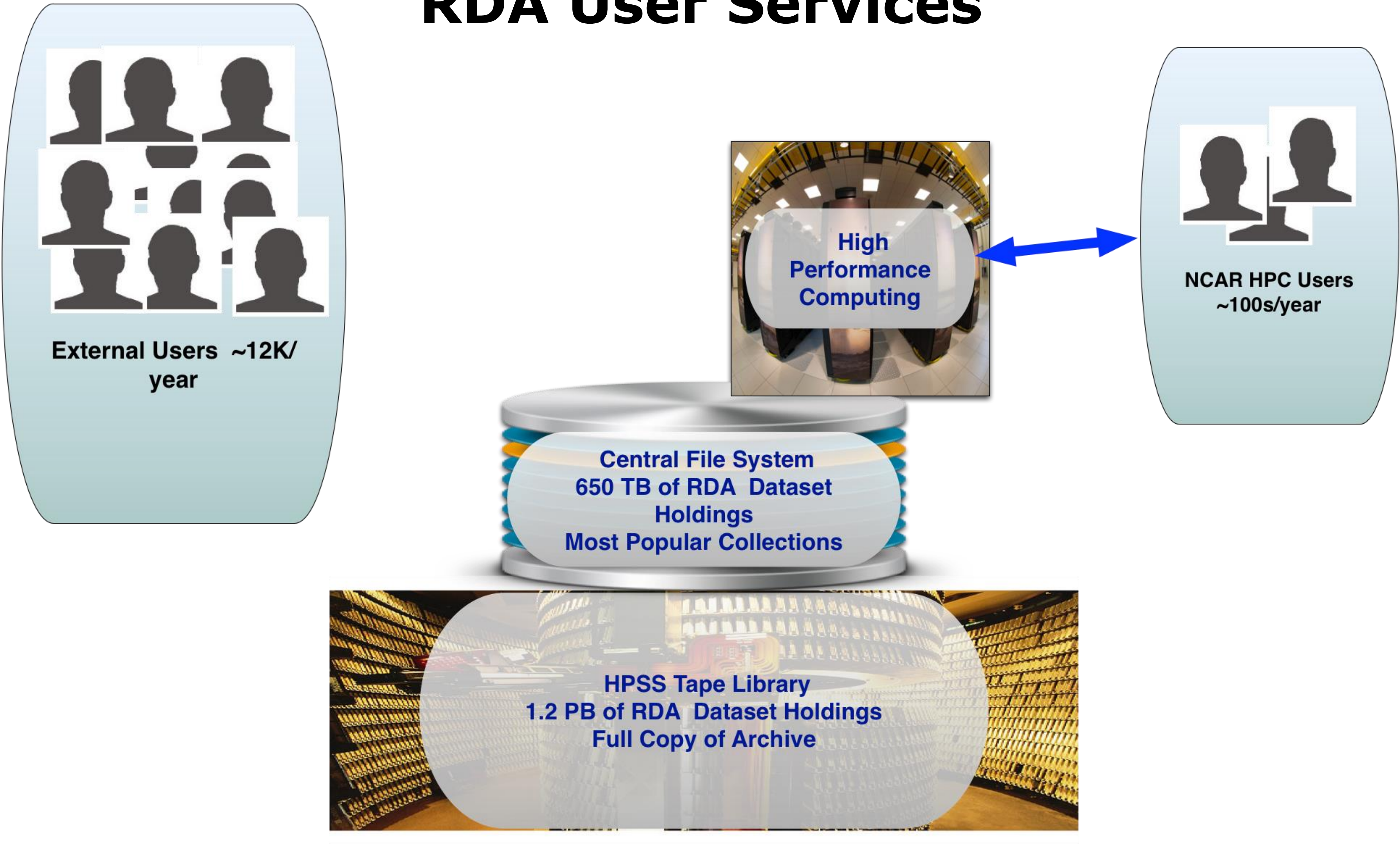
External Users ~12K/ year

NCAR HPC Users ~100s/year

Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

HPSS Tape Library
1.2 PB of RDA Dataset Holdings
Full Copy of Archive

# RDA User Services



External Users ~12K/ year

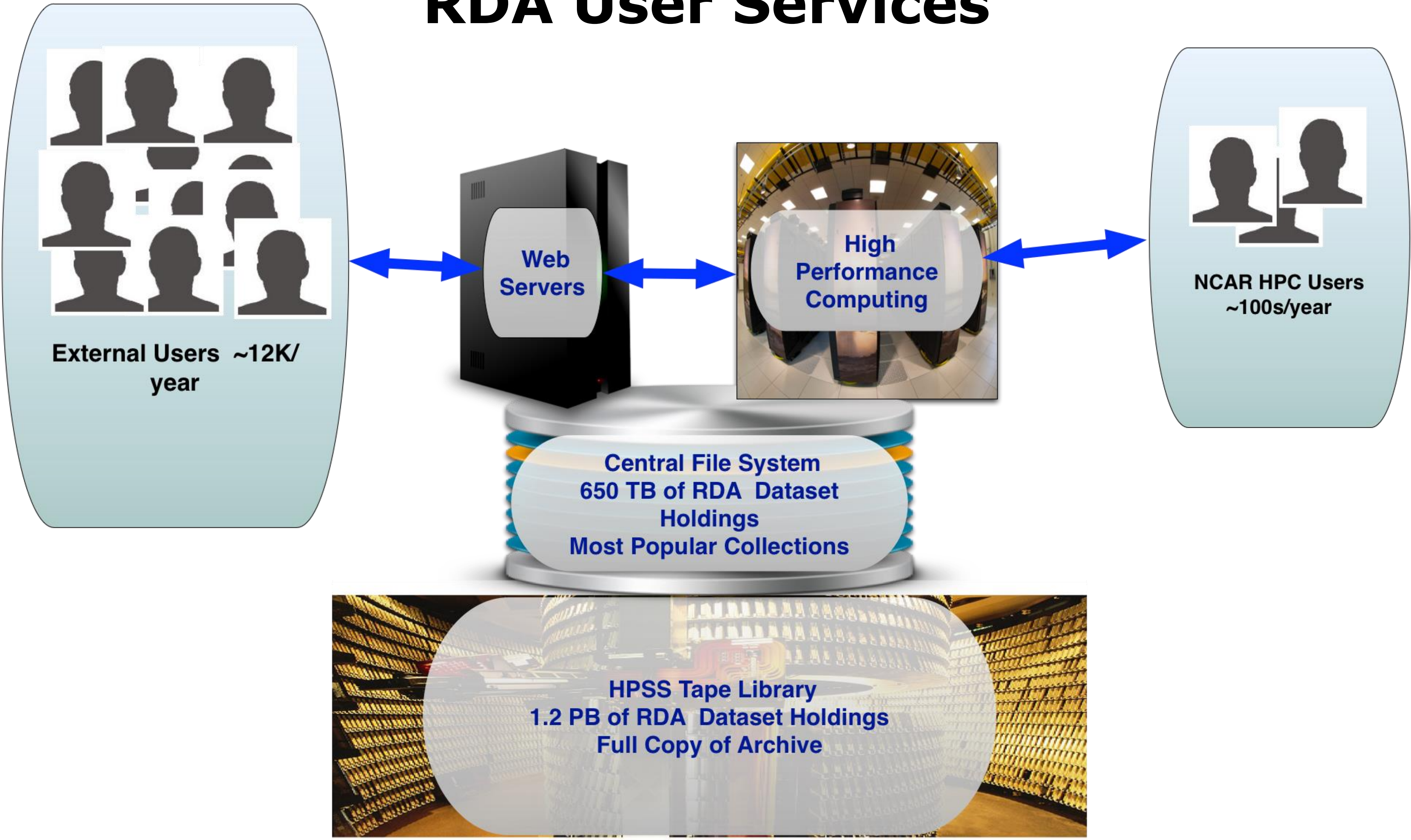
High Performance Computing

NCAR HPC Users ~100s/year

Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

HPSS Tape Library
1.2 PB of RDA Dataset Holdings
Full Copy of Archive

# RDA User Services



External Users ~12K/ year

Web Servers

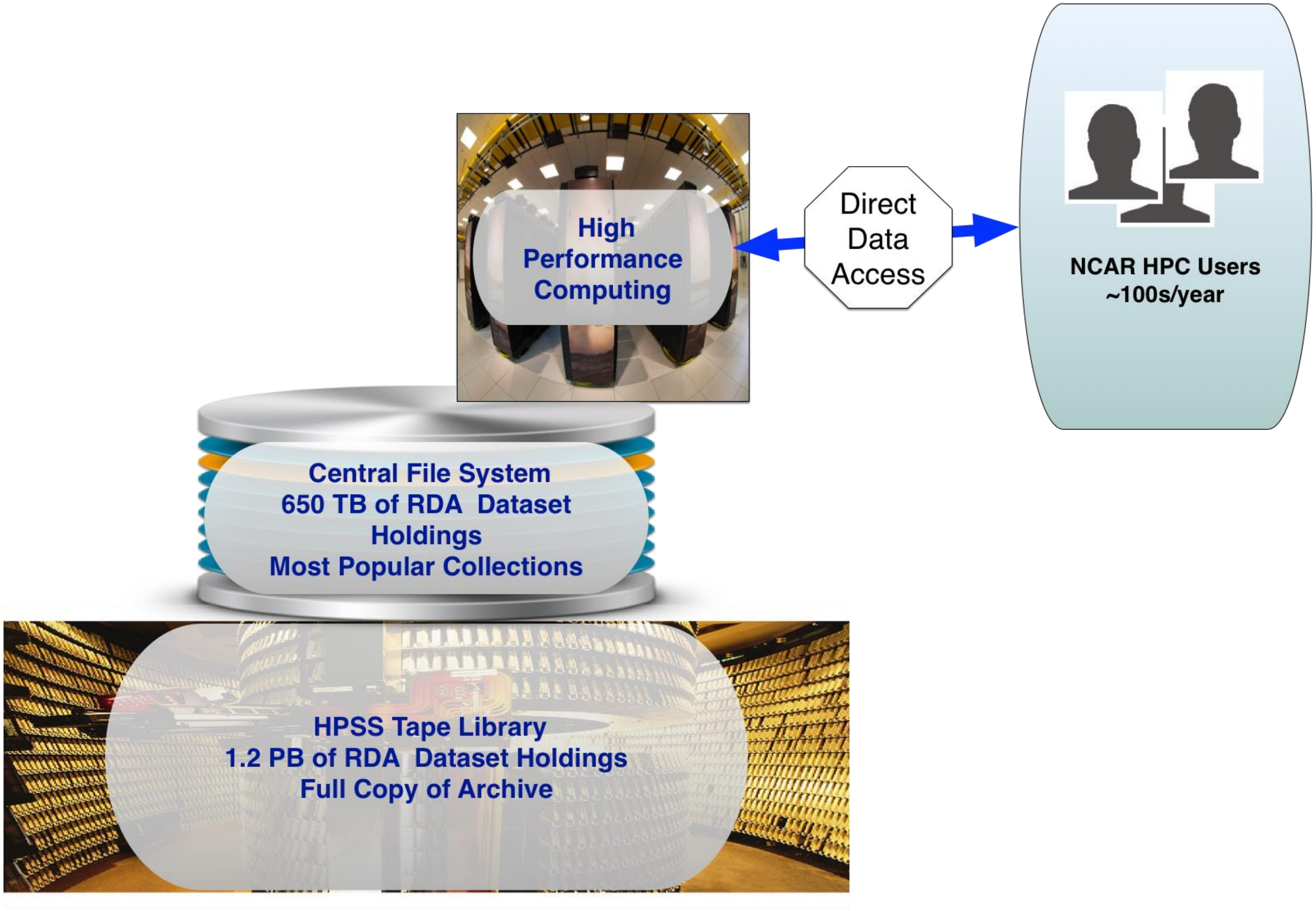
High Performance Computing

NCAR HPC Users ~100s/year

Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

HPSS Tape Library
1.2 PB of RDA Dataset Holdings
Full Copy of Archive

# RDA User Services



High Performance Computing

Direct Data Access

NCAR HPC Users ~100s/year

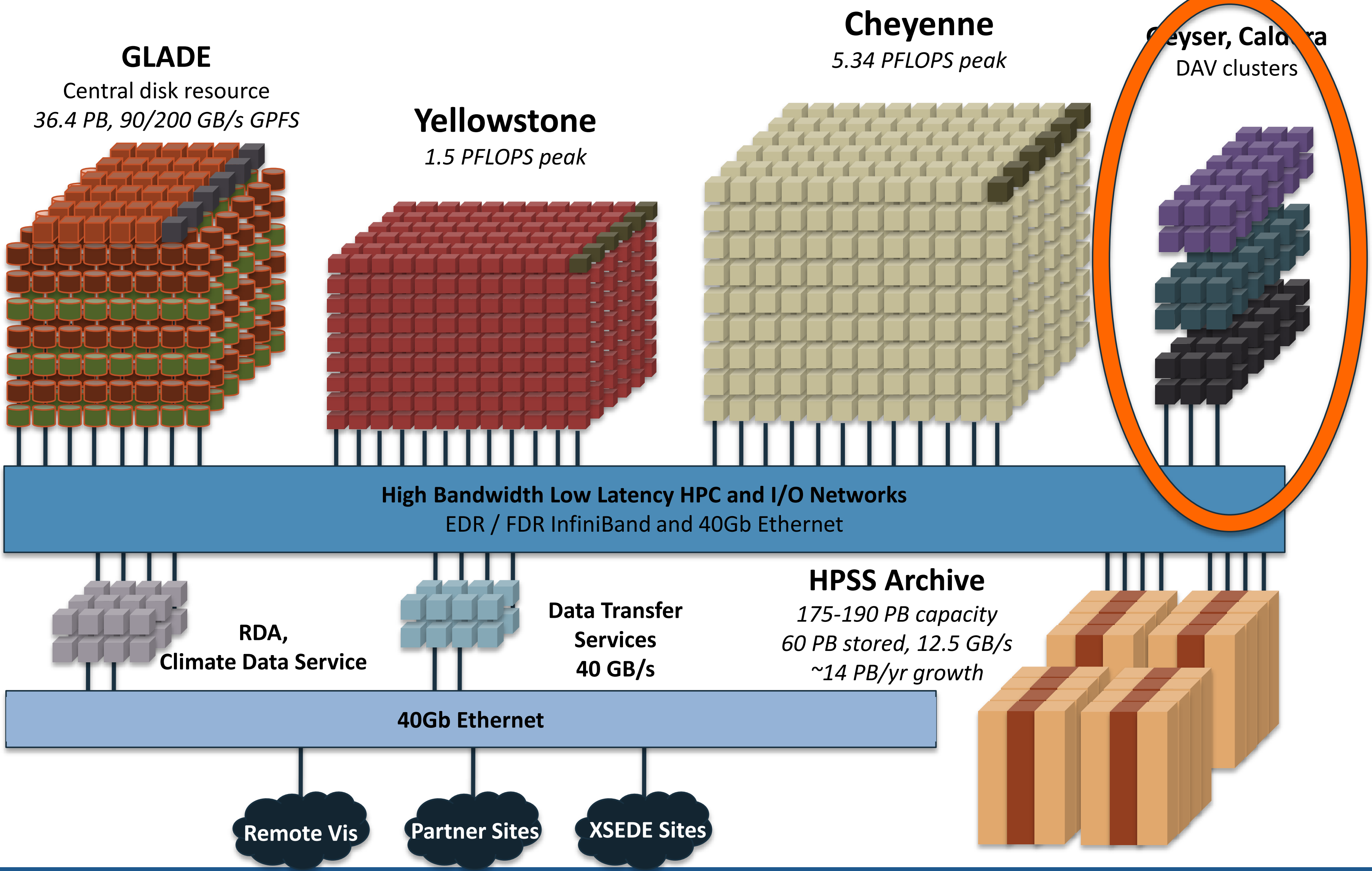
Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

HPSS Tape Library
1.2 PB of RDA Dataset Holdings
Full Copy of Archive

# NCAR Wyoming Supercomputing Center

- Opened in 2012
- Building performing great
- Ongoing improvements to achieve best operational efficiency
- Cheyenne (SGI) system brought online in Feb 2017 -#20 in top 500

# NCAR Data Intensive Supercomputing Environment



**GLADE**
Central disk resource
*36.4 PB, 90/200 GB/s GPFS*

**Yellowstone**
*1.5 PFLOPS peak*

**Cheyenne**
*5.34 PFLOPS peak*

**Geyser, Caldera**
DAV clusters

**High Bandwidth Low Latency HPC and I/O Networks**
EDR / FDR InfiniBand and 40Gb Ethernet

**RDA,**
**Climate Data Service**

**Data Transfer**
**Services**
**40 GB/s**

**HPSS Archive**
*175-190 PB capacity*
*60 PB stored, 12.5 GB/s*
*~14 PB/yr growth*

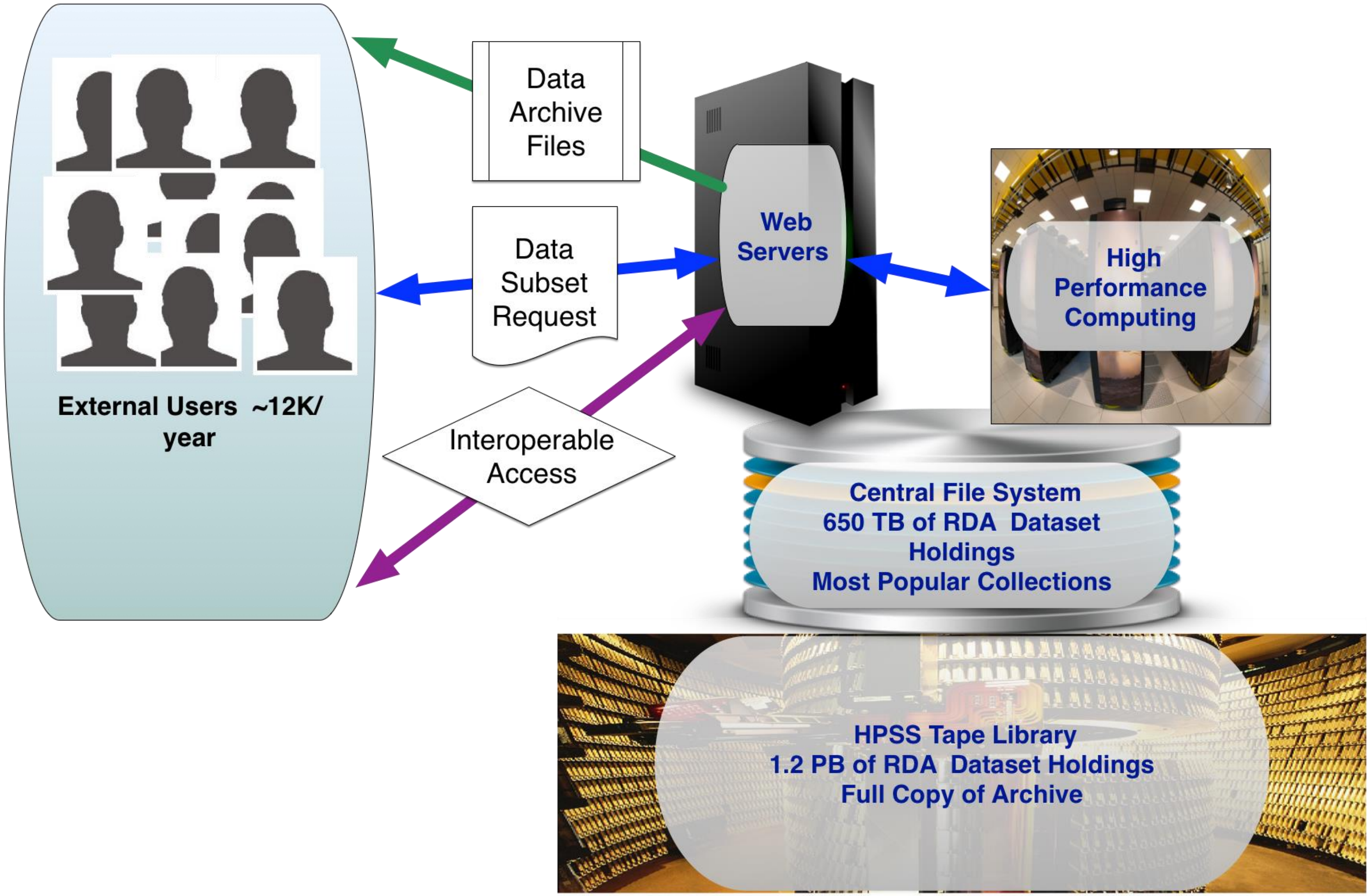
**40Gb Ethernet**

**Remote Vis**

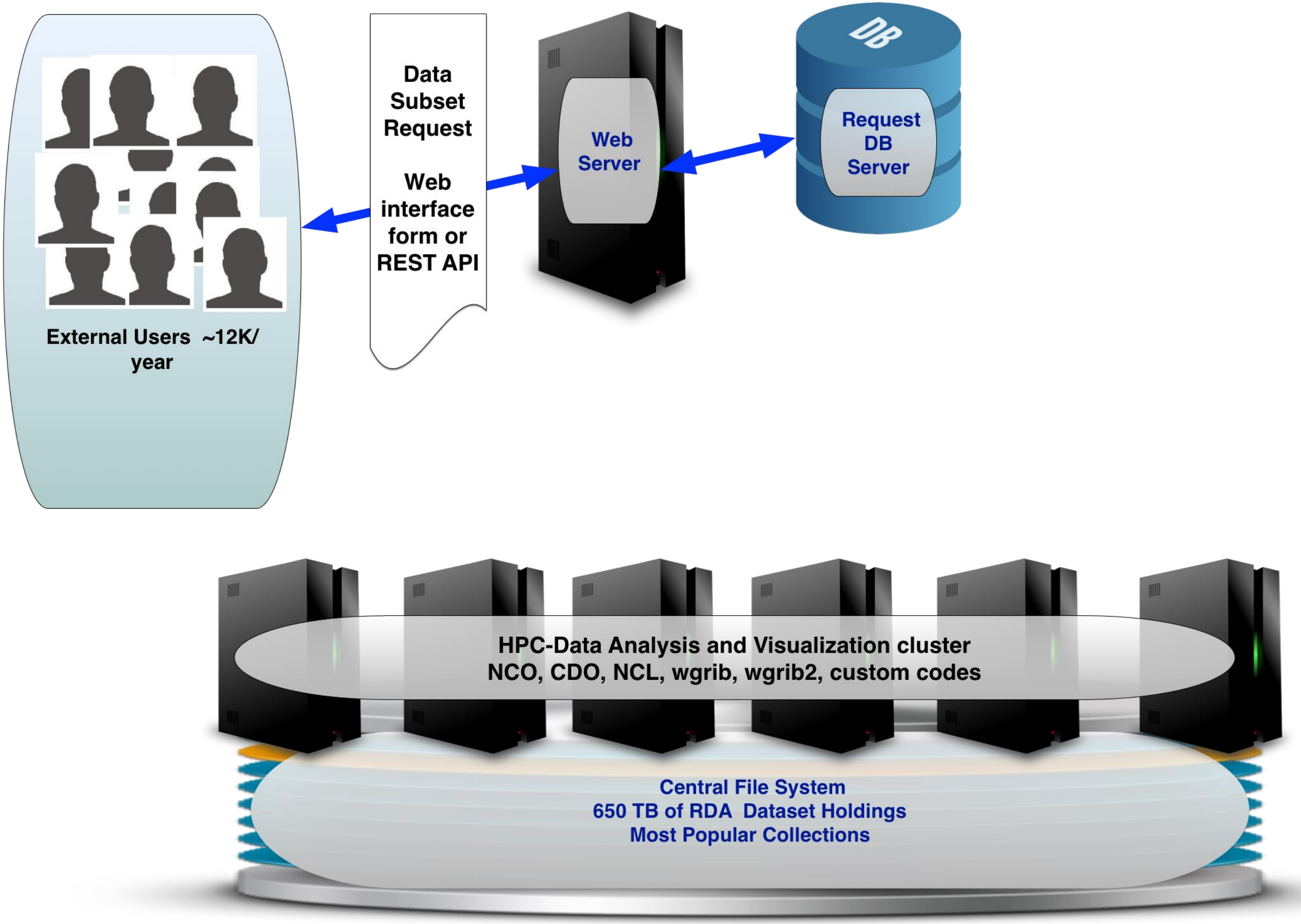
**Partner Sites**

**XSEDE Sites**

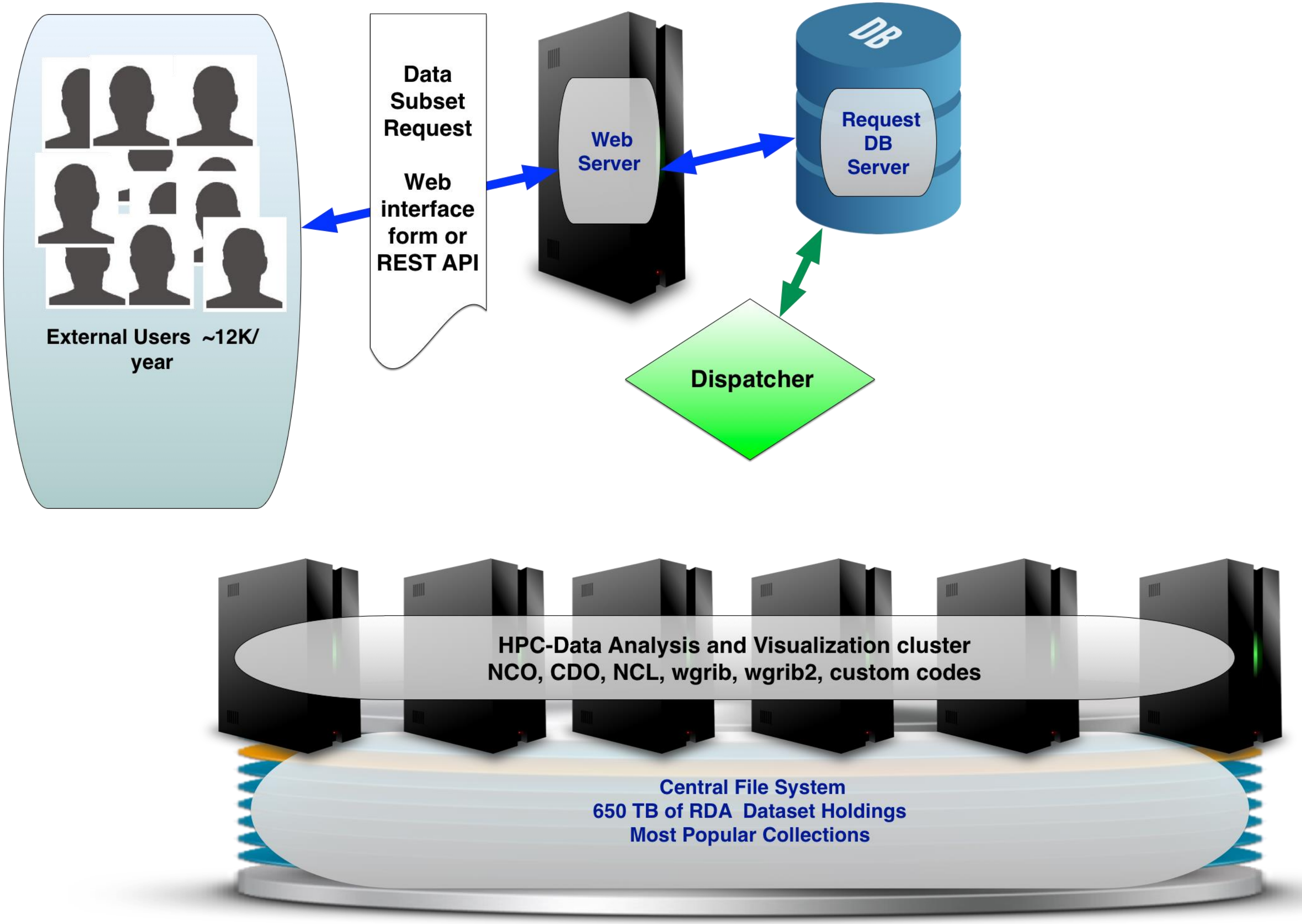
# RDA User Services



# RDA User Services
# Data Volume Reduction through HPC Processing


External Users ~12K/ year
Data Subset Request
Web interface form or REST API
Web Server
DB
Request DB Server
HPC-Data Analysis and Visualization cluster
NCO, CDO, NCL, wgrib, wgrib2, custom codes
Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

# RDA User Services
# Data Volume Reduction through HPC Processing



External Users ~12K/ year

Data Subset Request

Web interface form or REST API

Web Server

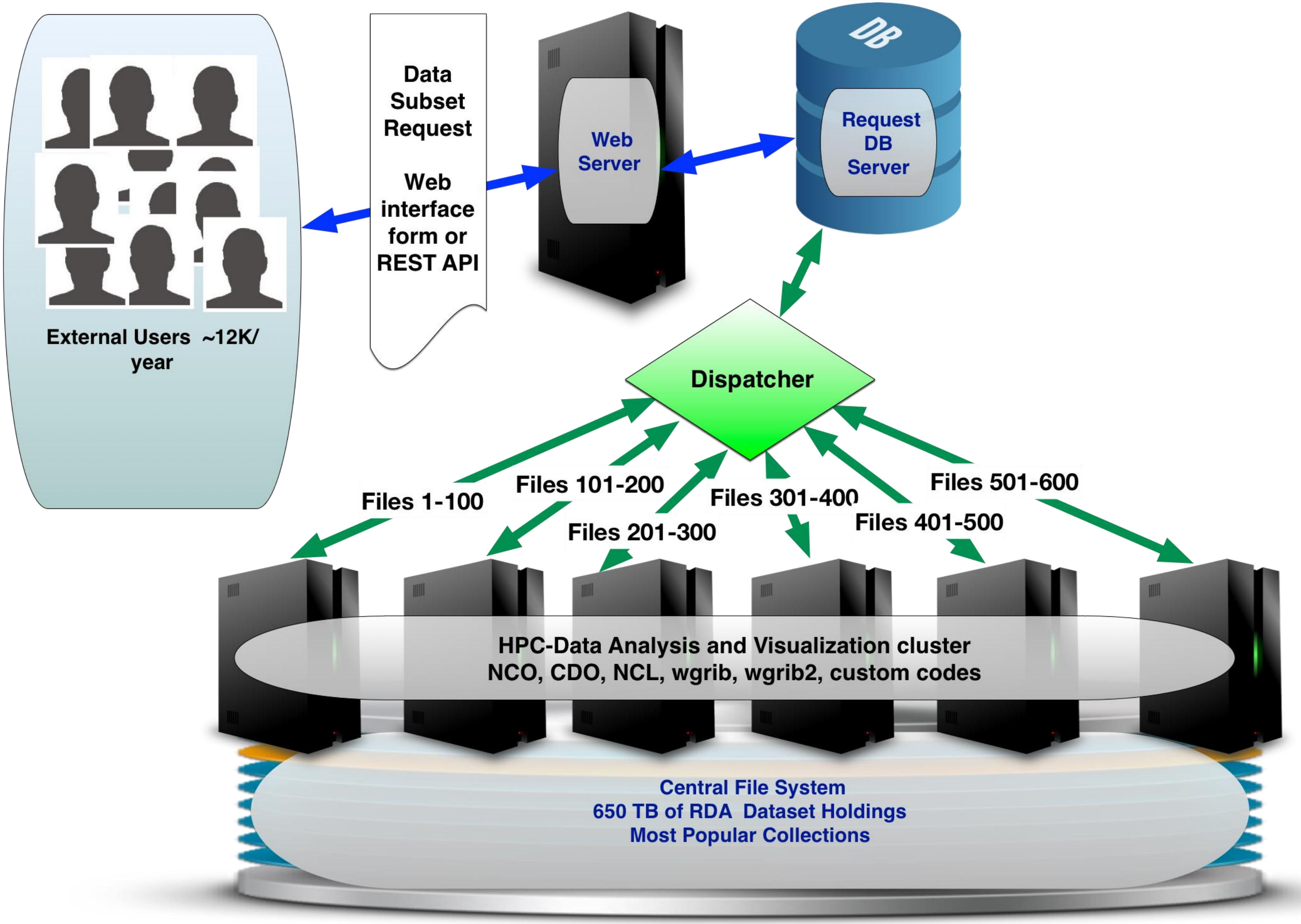
Request DB Server

Dispatcher

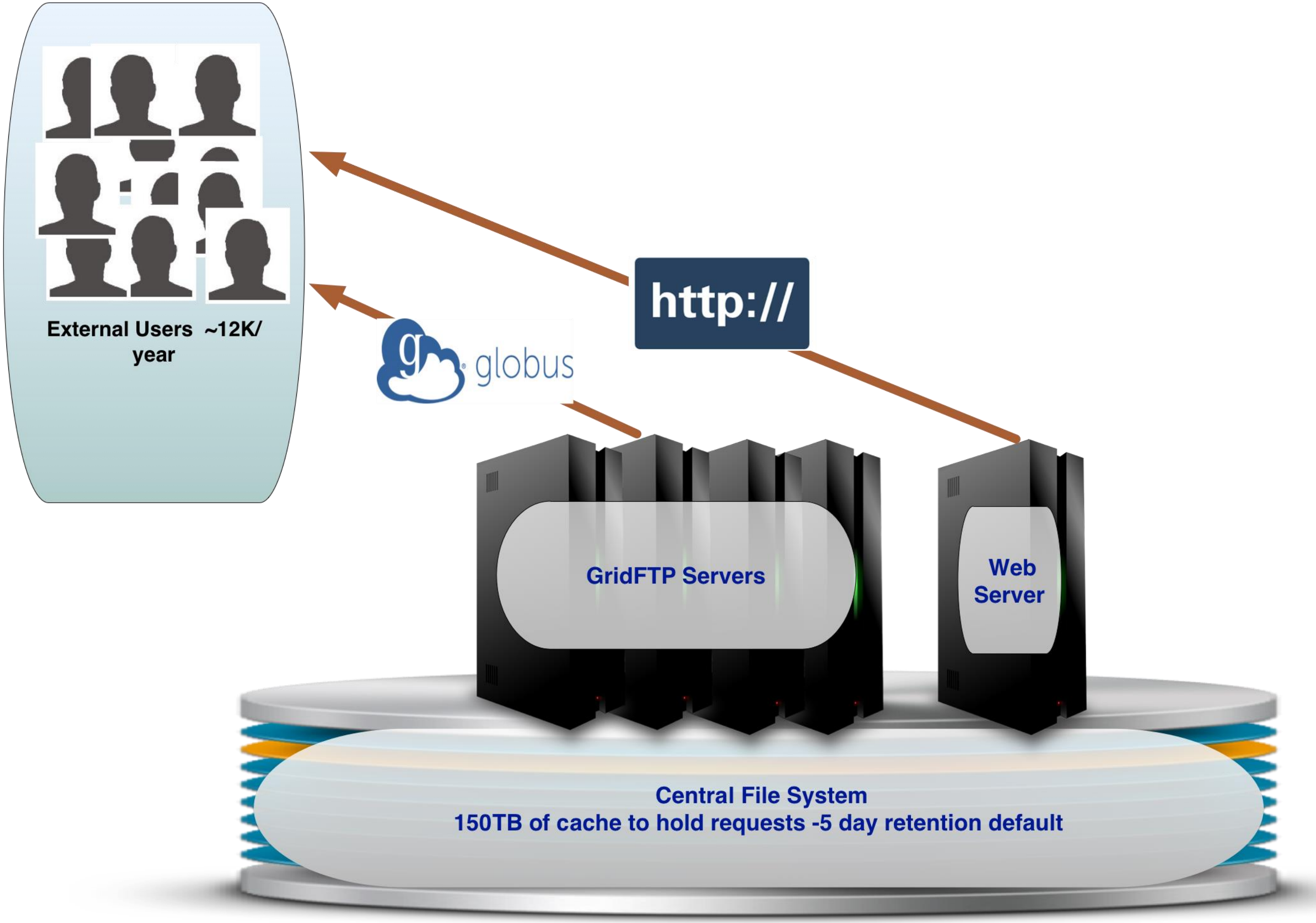
HPC-Data Analysis and Visualization cluster
NCO, CDO, NCL, wgrib, wgrib2, custom codes

Central File System
650 TB of RDA Dataset Holdings
Most Popular Collections

# RDA User Services
# Data Volume Reduction through HPC Processing

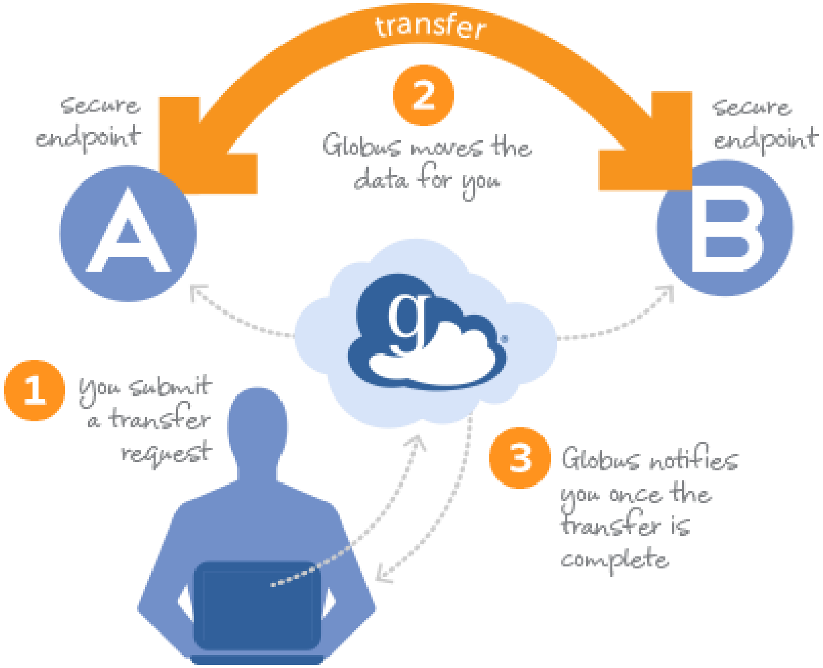


# RDA User Services
# Data Transfer Mechanisms

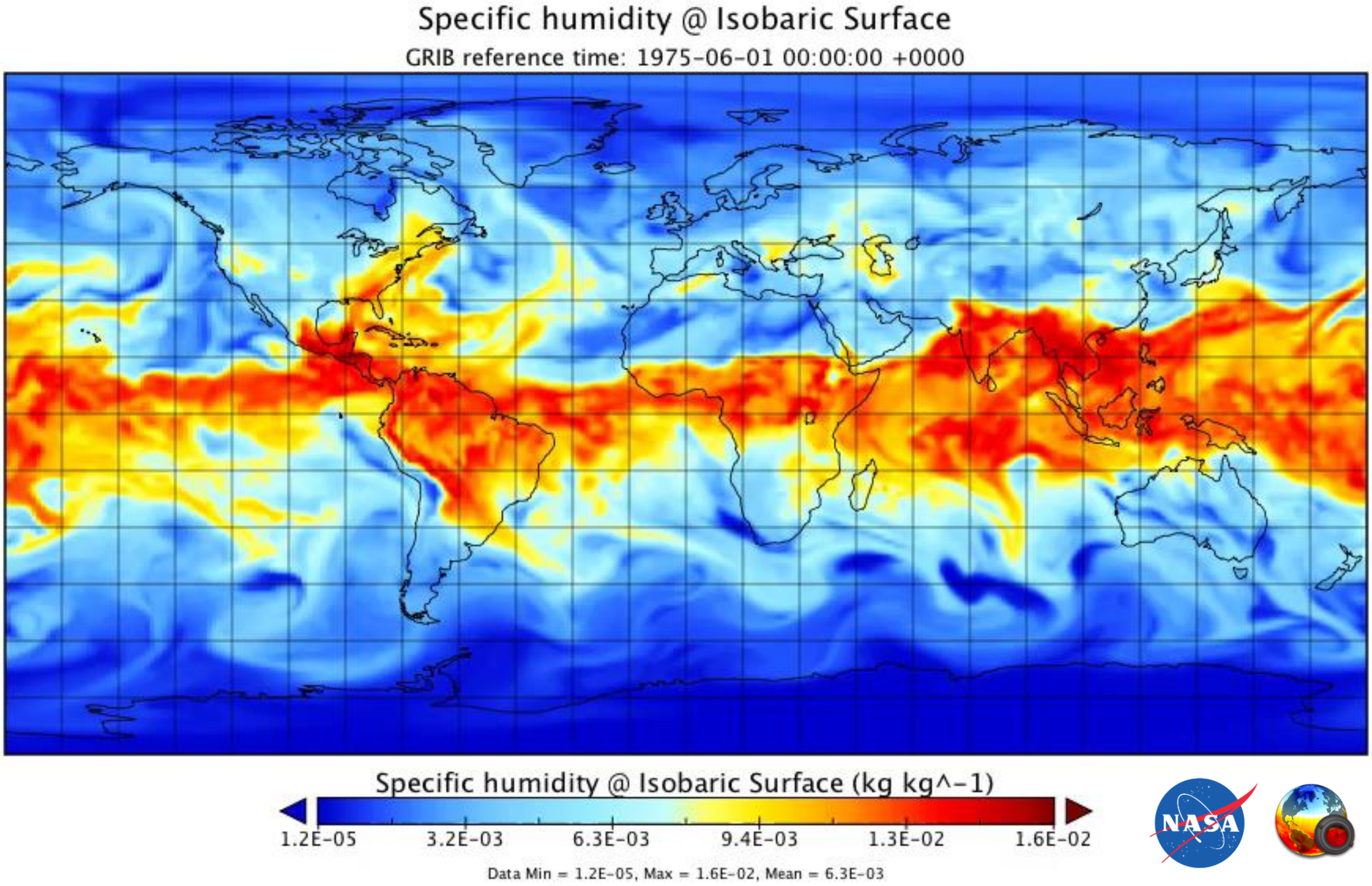

External Users ~12K/ year
http://
globus
GridFTP Servers
Web Server
Central File System
150TB of cache to hold requests -5 day retention default

# RDA User Services
# Data Transfer Mechanisms
# Globus Advantages

- **Reliable, secure, high-performance** file transfer
- "Fire and forget"
- Automatic fault recovery
- Powerful GUI, APIs, and CLI


transfer
secure endpoint
A
2
Globus moves the data for you
secure endpoint
B
1
You submit a transfer request
3
Globus notifies you once the transfer is complete

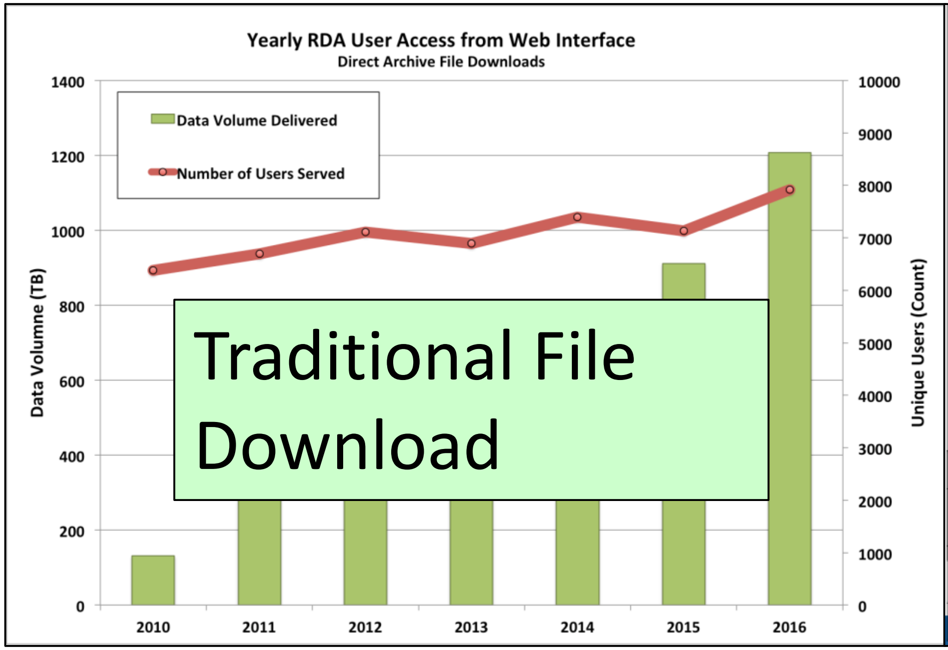
# RDA User Services
# OpenDAP Access Example Product, THREDDS


Specific humidity @ Isobaric Surface
GRIB reference time: 1975-06-01 00:00:00 +0000
Specific humidity @ Isobaric Surface (kg kg^-1)
1.2E-05
3.2E-03
6.3E-03
9.4E-03
1.3E-02
1.6E-02
Data Min = 1.2E-05, Max = 1.6E-02, Mean = 6.3E-03

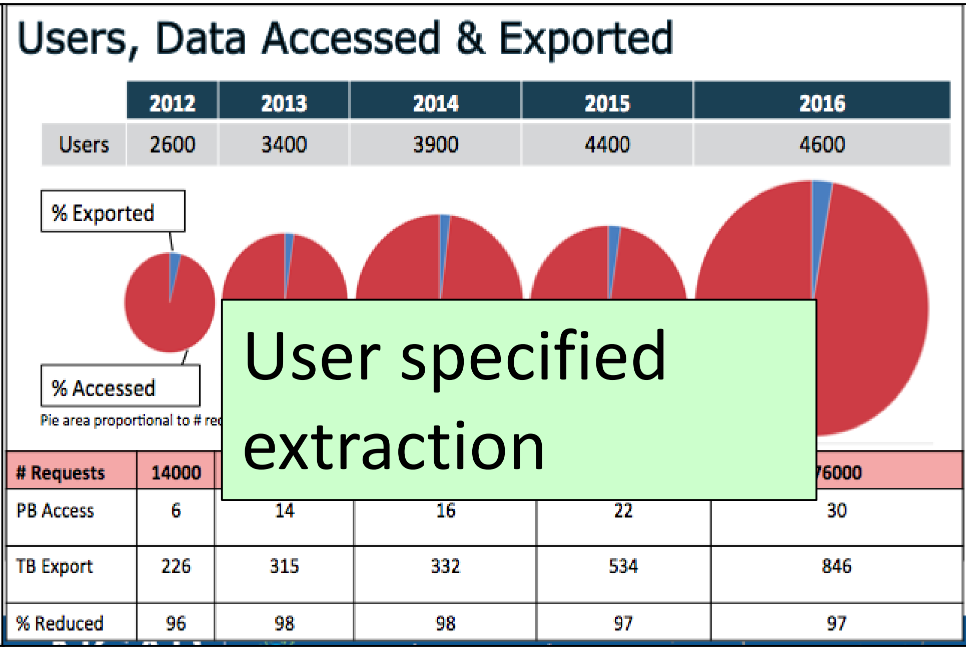
# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# Current Services Overview – Usage metrics



Traditional File Download



## Users, Data Accessed & Exported

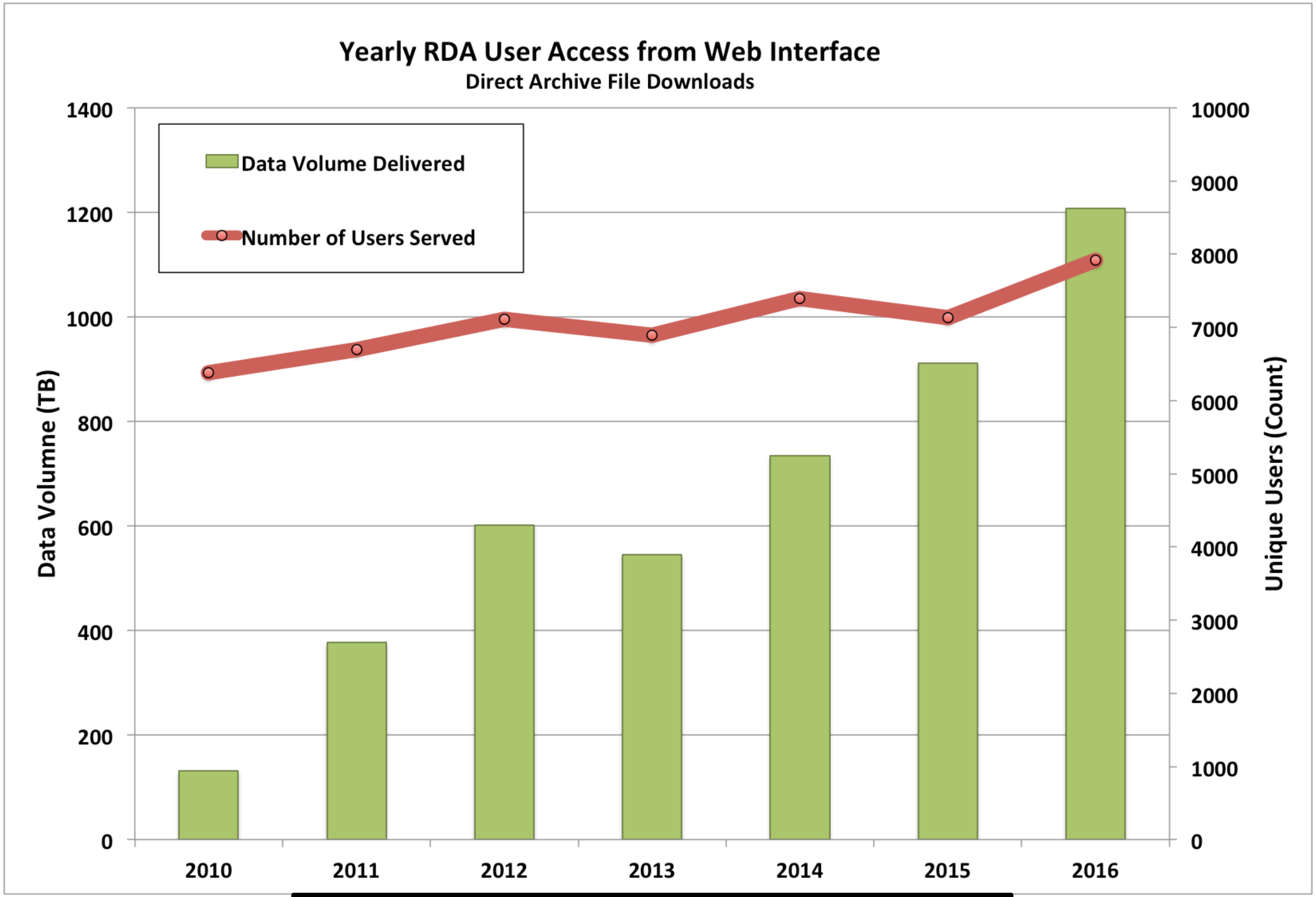
| | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|
| Users | 2600 | 3400 | 3900 | 4400 | 4600 |

% Exported
% Accessed
Pie area proportional to # re

| # Requests | 14000 | | | | 76000 |
|---|---|---|---|---|---|
| PB Access | 6 | 14 | 16 | 22 | 30 |
| TB Export | 226 | 315 | 332 | 534 | 846 |
| % Reduced | 96 | 98 | 98 | 97 | 97 |

User specified extraction

# Traditional File Download – Usage metrics

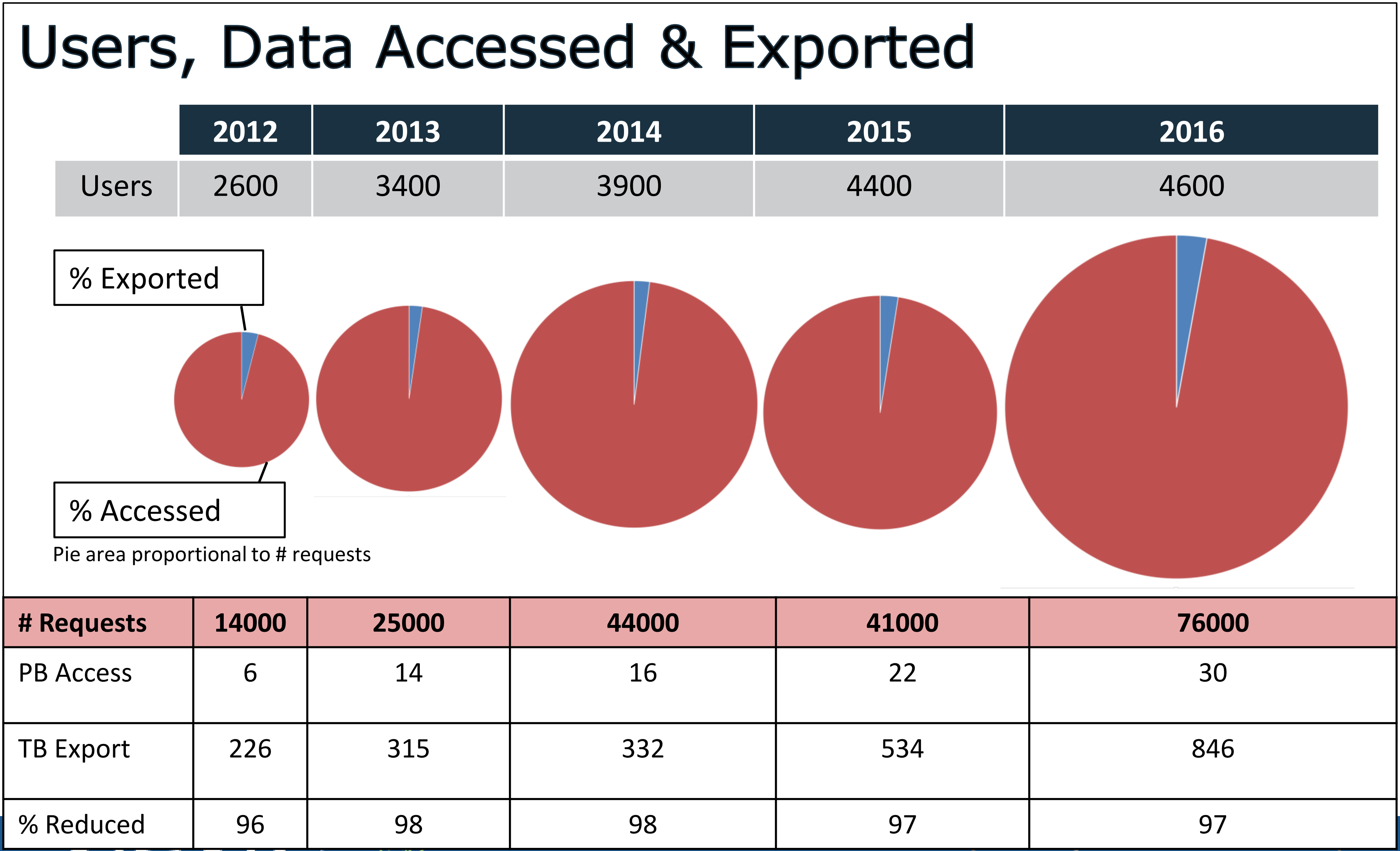


In FY 2016

- 7900 users
- 1.2 PB of data delivered

# Breaking the Big Data barrier for users

- Users select what they want, RDA extracts what they need



## Users, Data Accessed & Exported

| | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|
| Users | 2600 | 3400 | 3900 | 4400 | 4600 |

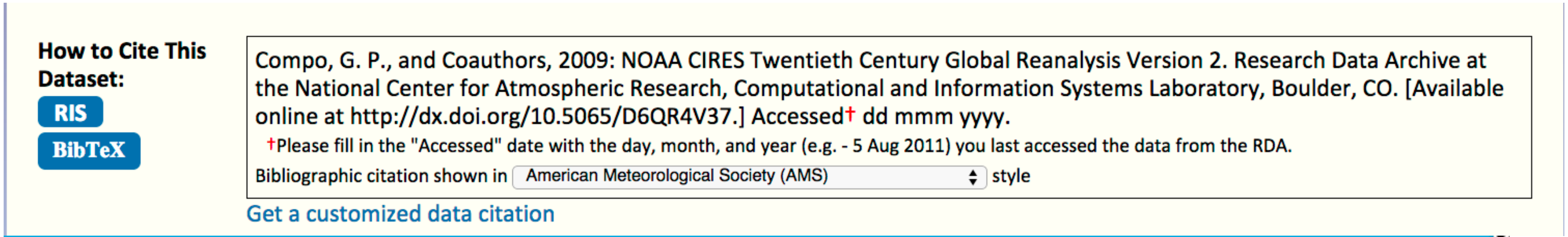
Pie area proportional to # requests

| # Requests | 14000 | 25000 | 44000 | 41000 | 76000 |
|---|---|---|---|---|---|
| PB Access | 6 | 14 | 16 | 22 | 30 |
| TB Export | 226 | 315 | 332 | 534 | 846 |
| % Reduced | 96 | 98 | 98 | 97 | 97 |

# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# User Outreach and Support
# -Server Generated Data Citation



**How to Cite This Dataset:**

RIS

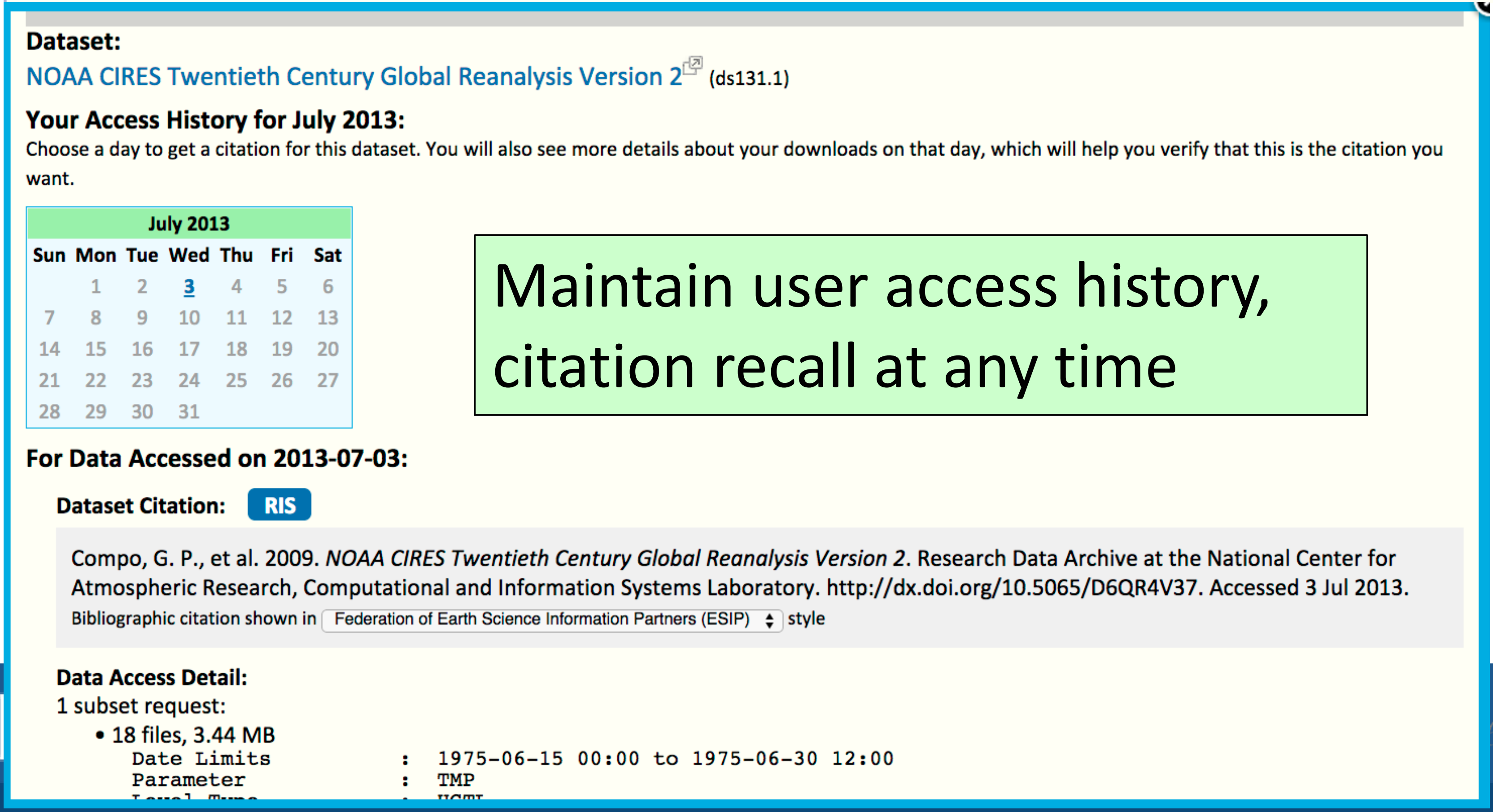
BibTeX

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 American Meteorological Society (AMS) style

Get a customized data citation



**Dataset:**

NOAA CIRES Twentieth Century Global Reanalysis Version 2 (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 | 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 | | | |

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 Federation of Earth Science Information Partners (ESIP) style

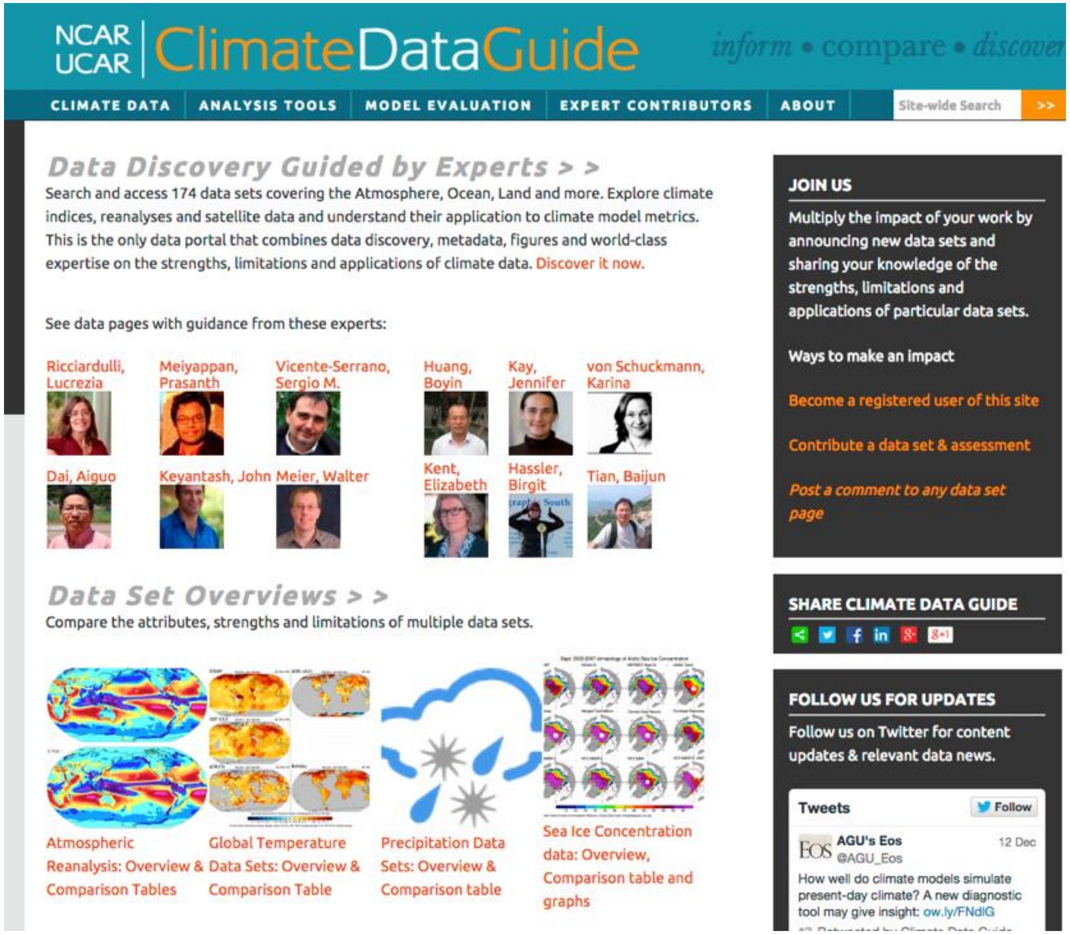
**Data Access Detail:**

1 subset request:

- 18 files, 3.44 MB

```
    Date Limits          :  1975-06-15 00:00 to 1975-06-30 12:00
    Parameter            :  TMP
```

# User Outreach and Support -Advise User of Best Dataset Collection(s) for purpose



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:

Ricciardulli, Lucrezia; Meiyappan, Prasanth; Vicente-Serrano, Sergio M.; Huang, Boyin; Kay, Jennifer; von Schuckmann, Karina; Dai, Aiguo; Keyantash, John; Meier, Walter; Kent, Elizabeth; Hassler, Birgit; Tian, Baijun

## Data Set Overviews > >

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

Atmospheric Reanalysis: Overview & Comparison Tables

Global Temperature Data Sets: Overview & 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**

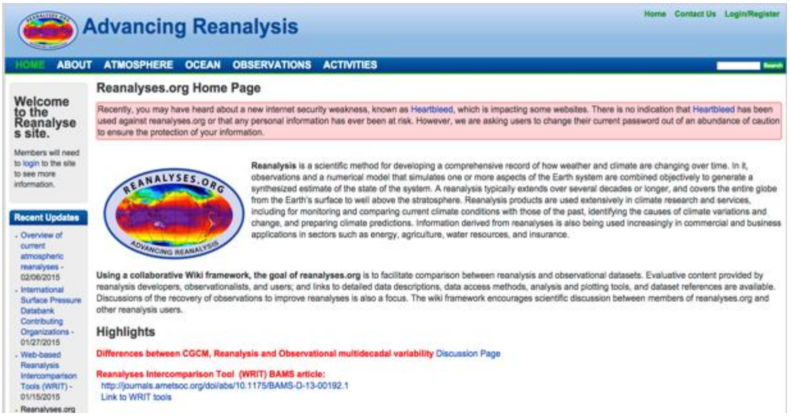
**FOLLOW US FOR UPDATES**

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

Tweets | Follow

AGU's Eos @AGU_Eos 12 Dec

How well do climate models simulate present-day climate? A new diagnostic tool may give insight: ow.ly/FNdlG


Advancing Reanalysis
Home Contact Us Login/Register
HOME ABOUT ATMOSPHERE OCEAN OBSERVATIONS ACTIVITIES
Search
REANALYSES.ORG
ADVANCING REANALYSIS

**Welcome to the Reanalyses site.**

Members will need to login to the site to see more information.

**Recent Updates**

- Overview of current atmospheric reanalyses - 02/06/2015
- International Surface Pressure Databank Contributing Organizations - 01/27/2015
- Web-based Reanalysis Intercomparison Tools (WRIT) - 01/15/2015
- Reanalyses.org

### Reanalyses.org Home Page

Recently, you may have heard about a new internet security weakness, known as Heartbleed, which is impacting some websites. There is no indication that Heartbleed has been used against reanalyses.org or that any personal information has ever been at risk. However, we are asking users to change their current password out of an abundance of caution to ensure the protection of your information.

**Reanalysis** is a scientific method for developing a comprehensive record of how weather and climate are changing over time. In it, observations and a numerical model that simulates one or more aspects of the Earth system are combined objectively to generate a synthesized estimate of the state of the system. A reanalysis typically extends over several decades or longer, and covers the entire globe from the Earth's surface to well above the stratosphere. Reanalysis products are used extensively in climate research and services, including for monitoring and comparing current climate conditions with those of the past, identifying the causes of climate variations and change, and preparing climate predictions. Information derived from reanalyses is also being used increasingly in commercial and business applications in sectors such as energy, agriculture, water resources, and insurance.

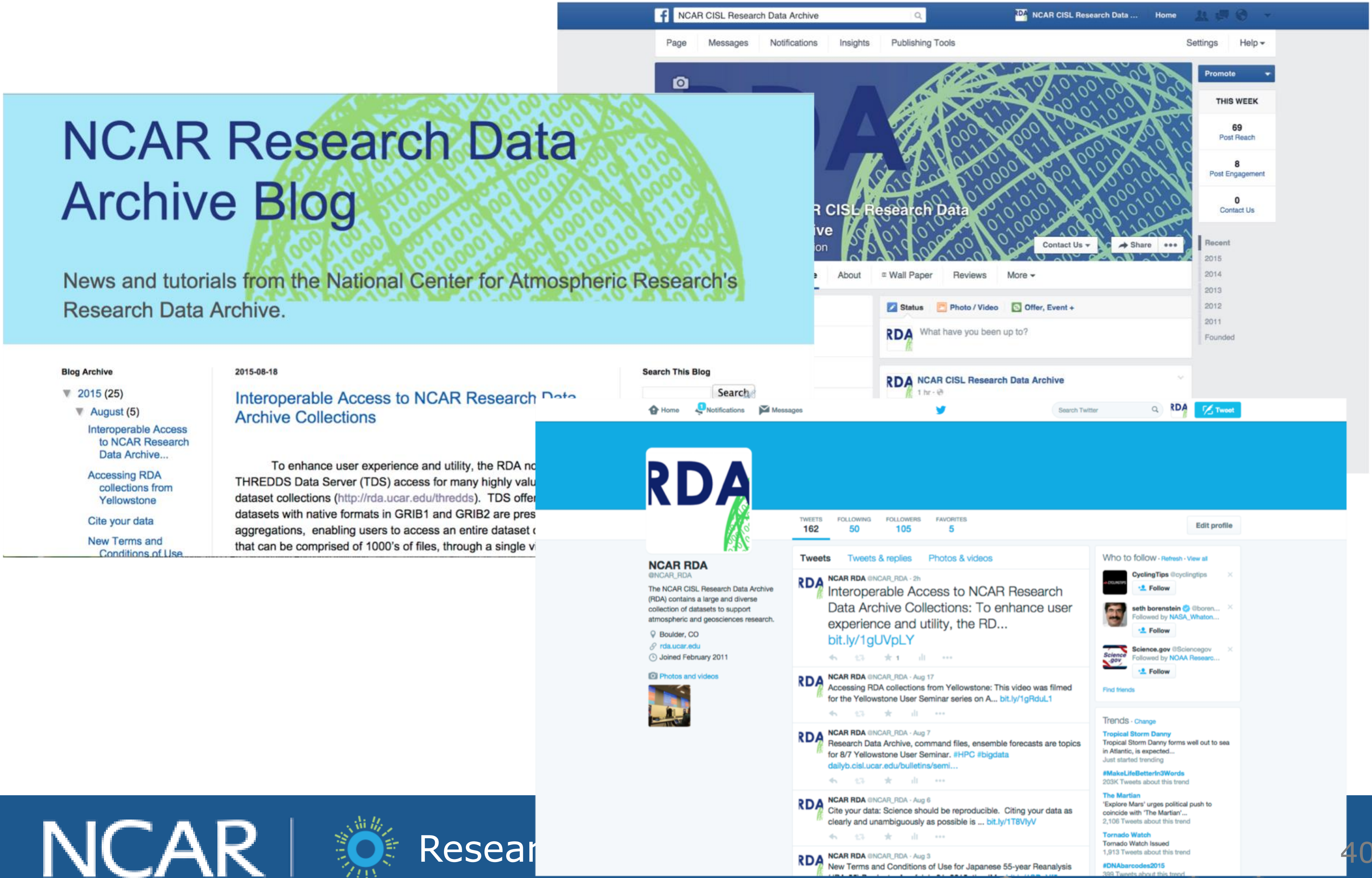
**Using a collaborative Wiki framework, the goal of reanalyses.org** is to facilitate comparison between reanalysis and observational datasets. Evaluative content provided by reanalysis developers, observationalists, and users; and links to detailed data descriptions, data access methods, analysis and plotting tools, and dataset references are available. Discussions of the recovery of observations to improve reanalyses is also a focus. The wiki framework encourages scientific discussion between members of reanalyses.org and other reanalysis users.

### Highlights

**Differences between CGCM, Reanalysis and Observational multidecadal variability** Discussion Page

**Reanalyses Intercomparison Tool (WRIT) BAMS article:**
http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-13-00192.1
Link to WRIT tools

# User Outreach and Support
# Social Media, with User Video Tutorials



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 no THREDDS Data Server (TDS) access for many highly valu dataset collections (http://rda.ucar.edu/thredds). TDS offe datasets with native formats in GRIB1 and GRIB2 are pres aggregations, enabling users to access an entire dataset that can be comprised of 1000's of files, through a single v

Search This Blog

R CISL Research Data
ive
ion

THIS WEEK

69 Post Reach

8 Post Engagement

0 Contact Us

NCAR CISL Research Data Archive
1 hr

| TWEETS | FOLLOWING | FOLLOWERS | FAVORITES |
|---|---|---|---|
| 162 | 50 | 105 | 5 |

NCAR RDA
@NCAR_RDA

The NCAR CISL Research Data Archive (RDA) contains a large and diverse collection of datasets to support atmospheric and geosciences research.

Boulder, CO
rda.ucar.edu
Joined February 2011

Photos and videos

NCAR RDA @NCAR_RDA · 2h
Interoperable Access to NCAR Research Data Archive Collections: To enhance user experience and utility, the RD... bit.ly/1gUVpLY

NCAR RDA @NCAR_RDA · Aug 17
Accessing RDA collections from Yellowstone: This video was filmed for the Yellowstone User Seminar series on A... bit.ly/1gRduL1

NCAR RDA @NCAR_RDA · Aug 7
Research Data Archive, command files, ensemble forecasts are topics for 8/7 Yellowstone User Seminar. #HPC #bigdata dailyb.cisl.ucar.edu/bulletins/semi...

NCAR RDA @NCAR_RDA · Aug 6
Cite your data: Science should be reproducible. Citing your data as clearly and unambiguously as possible is ... bit.ly/1T8VlyV

NCAR RDA @NCAR_RDA · Aug 3
New Terms and Conditions of Use for Japanese 55-year Reanalysis

Who to follow · Refresh · View all

CyclingTips @cyclingtips
Follow

seth borenstein @boren...
Followed by NASA_Whaton...
Follow

Science.gov @Sciencegov
Followed by NOAA Researc...
Follow

Find friends

Trends · Change

Tropical Storm Danny
Tropical Storm Danny forms well out to sea in Atlantic, is expected...
Just started trending

#MakeLifeBetterIn3Words
203K Tweets about this trend

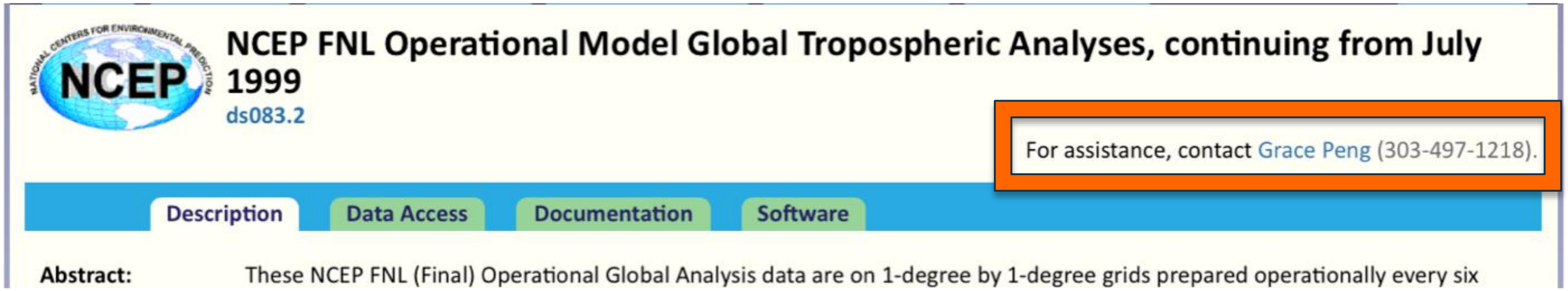
The Martian
'Explore Mars' urges political push to coincide with 'The Martian'...
2,106 Tweets about this trend

Tornado Watch
Tornado Watch Issued
1,913 Tweets about this trend

#DNAbarcodes2015
399 Tweets about this trend

# User Outreach and Support
# Direct Help

- Dataset consultant listed on each dataset homepage



NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999

ds083.2

For assistance, contact Grace Peng (303-497-1218).

Abstract: 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

# Highlights

- Overview of NCAR and the RDA
- Data Management Strategies
- Metadata Overview
- RDA User Services
- Usage Metrics
- User Outreach and User Support
- Lessons Learned and Future Directions

# 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, and data integrity validation upon ingest.
- Essential to co-locate data with HPC on disk

# Future Directions

- CAPSTONE Data Analysis Environment
  - Develop and share data analysis workflows
- SSD Storage to drive fast access
- Collaborate with NCAR GIS program
- RDA-ODB instance –Support NCAR DA research
- Distributed Ocean Matchup Service (DOMS)
  - Satellite to in-situ matchup distributed archives
- Continue to add high value assets to archive
  - ERA-5, ERA-CLIM
  - ASR
  - NCAR Reanalysis Products

Questions?

http://rda.ucar.edu

Doug Schuster
schuster@ucar.edu