User Driven Automatic Data Request Service

Providing User Access to TB-sized Datasets

Zaihua Ji
Steven Worley

Computational and Information Systems Laboratory
National Center for Atmospheric Research

http://rda.ucar.edu

NCAR SEA Conference 2013

Presentation Outline

- Introduction
- Research Data Archive (RDA) components
- Challenge of access to TB-sized datasets
- Design of DSRQST (DataSet ReQueST controller)
- Implementation of DSRQST
- Example
- Conclusion

Introduction

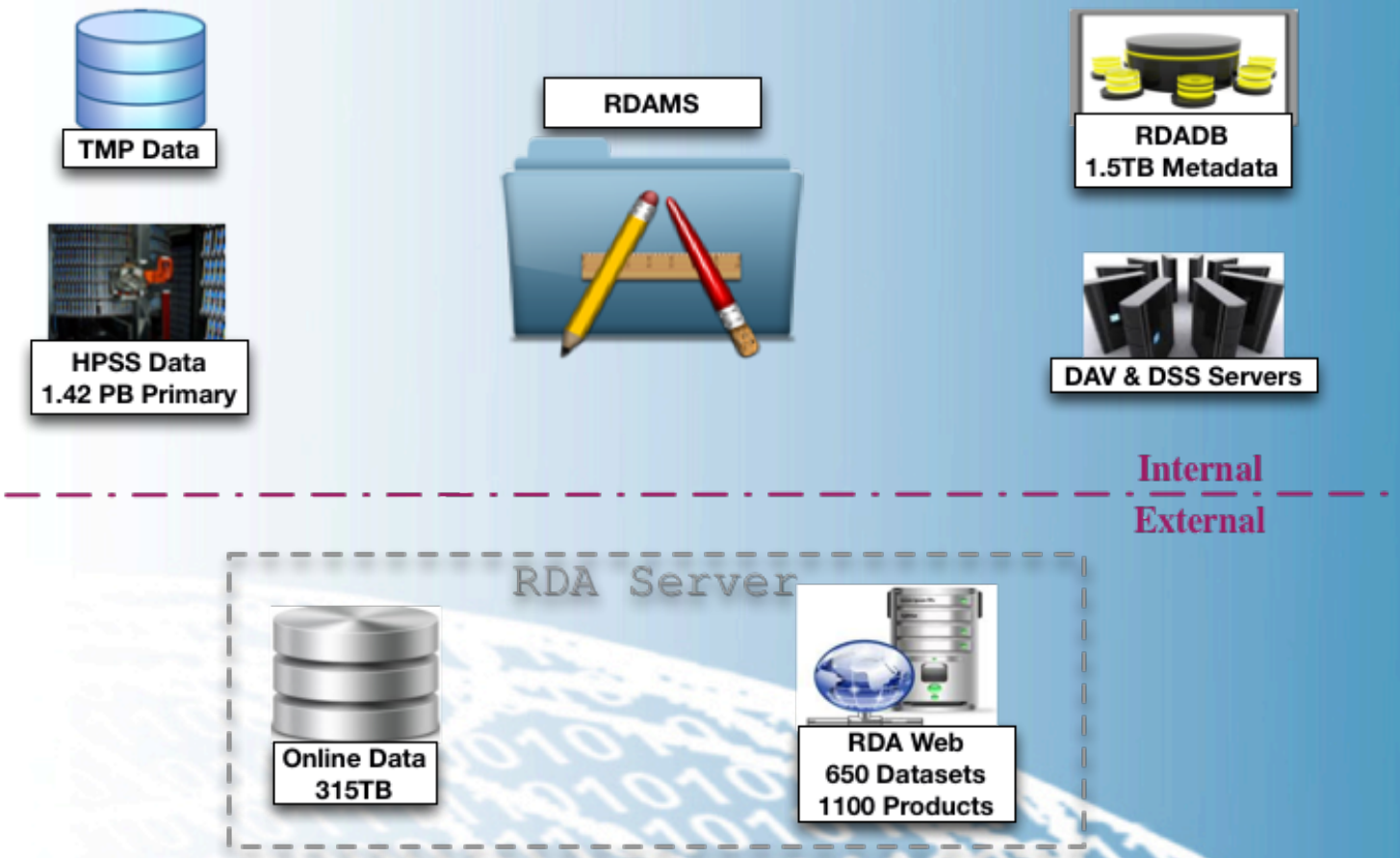
- Traditionally, data is delivered in fixed format and built in advance
- Presently, most data is delivered using network transfers
  - Direct file access from a shared disk system (GLADE)
  - Stage to disk from long-term archive (HPSS)
  - Supplemented with subsetting, and format conversion
- Past data service technologies did not scale
  - Dataset updates and access methods were developed independently
  - Not fault tolerant
    - Action driven by user GUIs not always robust
    - Data acquisition failures if data providers not timely and consistent
    - Data processing for users not tolerant of NCAR system down-time
  - Inaccurate user metrics collection
  - Archiving inconsistencies (metadata, file linkages, etc.)
  - Excess SE work time to resolve problems

Introduction - continued

New paradigm and scalable solution

- DSRQST - DataSet ReQueST controller/tool
  - Uses open source databases (MySQL) and locally written utilities
  - Records user requests and informs RDA specialists
  - Schedules processes to extract the data requested
  - Stages requested data to disk (remove duplication)
  - Informs users and specialists about data readiness
  - Purges the requested data and records activity metrics
- DSRQST also supports addition of new type of service

Research Data Archive Components



Research Data Archive Components - Internal

- TMP Data – Temporary storage for data processing

- RDAMS - RDA Management System
  - Retrieves data files from providers (remote)
  - Builds local data files (designed by specialist)
  - Archives data to online disk (external) and/or HPSS
  - Harvests file content standard metadata
  - Builds and stages data for user requests

Research Data Archive Components - Internal

- RDADB – RDA Database
  - File names, formats, and storage locations
  - Dataset discovery and file content metadata (1.5TB)
  - Data request information

- DAV & DSS Servers – Computer systems to process data
  - LSF batch processes on Yellowstone/Geysers
  - Interactive/direct processes on DSS systems

Research Data Archive Components - Internal

- HPSS Data – data on the NCAR High Performance Storage System
  - Primary archives of data (1.4PB)
  - Direct access for users with NCAR accounts
  - Indirect access for the public web users
  - Backup copies for disaster recovery (449TB)

CISL Computational & Information Systems Laboratory

NCAR

Research Data Archive Components - external

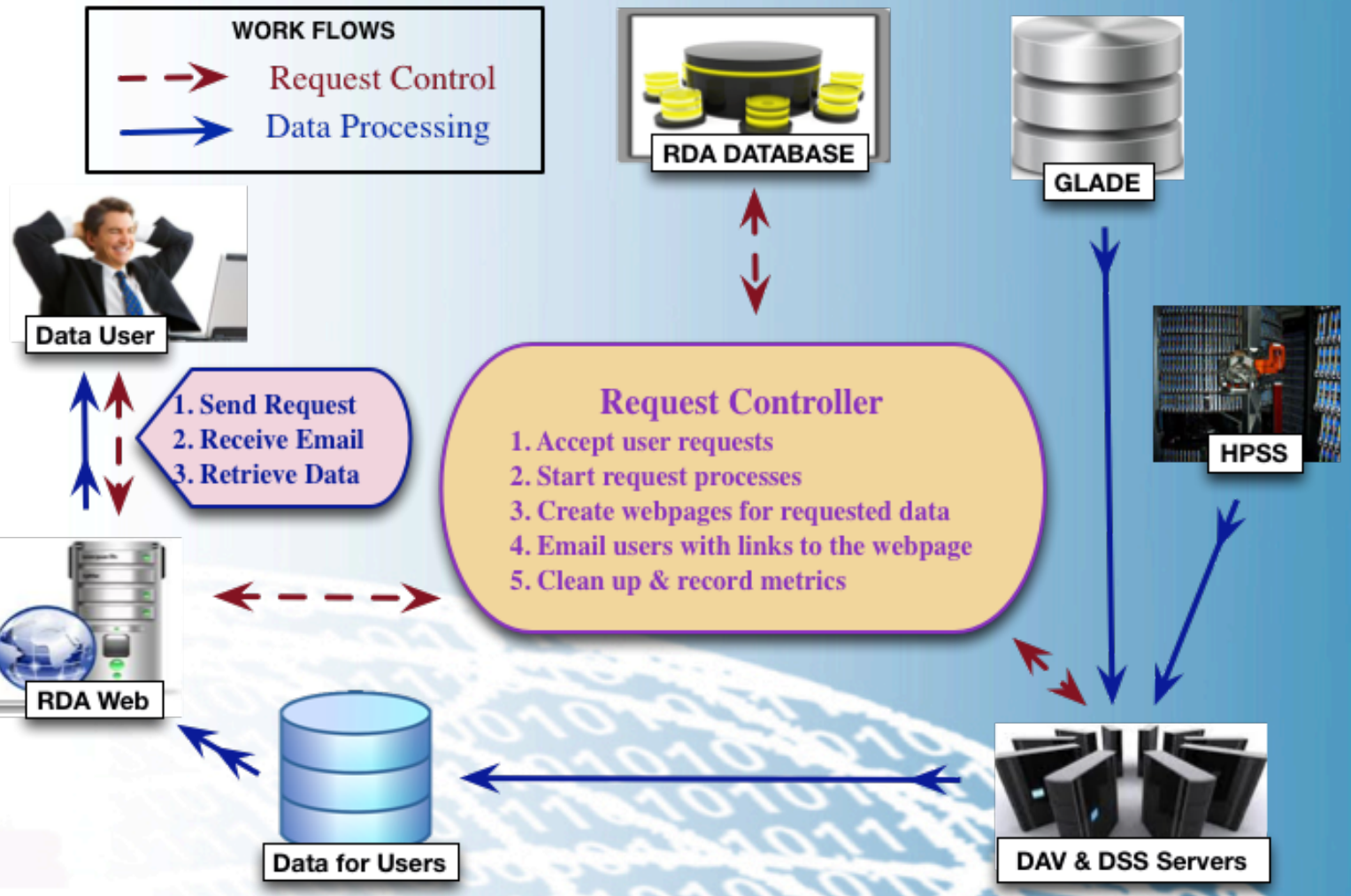
- RDA Web Interface – Open to public users
  - Access to 650 datasets and 1100 products
  - View and discover datasets
  - Download Online Data - real-time (315TB)
  - View Long-term archives on HPSS
  - Accept user data requests
  - Download data staged from one time user requests

- Online Data – Data on disk (GLADE)
  - Available through RDA Web Interface
  - Data files for direct download (315TB)
  - Directly readable on all CISL computers (need NCAR computing account)
  - Data files staged temporarily from one time user requests (20TB)

Challenges of Data Services for Large Datasets

- Variety of user demands
  - Direct file accesses
  - Stage to disk from long-term archived data
  - Format conversions
  - Subsetting
  - Others, such as Plotting, and Data Evaluation?
- Utilization of computing and storage resources
  - Diversity of computer configurations (LSF, etc.)
  - Frequent user requests
  - Large amount of data for a single request
  - Share the same data files for multiple requests
  - Limited time windows to preserve staged output data
- Other challenges
  - Restart interrupted data processing due to system outages
  - Scalability to including new services

Design of DSRQST – Data Request Service



Implementation of DSRQST

Three levels of programming configurations:

- Request Control – defines how a dataset product can be requested
- Request Information – records individual request constraints
- File Information –data file details for the request

Implementation of DSRQST – Request Control Configuration

- Control ID – Unique ID for a request control configuration
- Product ID – Identify a product, one or multiple products in a dataset
- Request Type – Types of requests, e.g. S for Subset
- Control Option – A for automatic, S for specialist to grant permission
- Command – Processing module to handle specific data diversity
- URL – Link to web interface that initiates the user request
- Compression – Compress for the output data, e.g. GZ
- Valid Period – Number of days to preserve the output for the user
- Host Name – Specify one or multiple computer nodes to process the request

Implementation of DSRQST – Request Information Configuration

- Request ID – Unique ID for the request
- User Email – User email address
- Product ID – Product Identification, one or multiple products in a dataset
- Request Type – Types of requests, e.g. S for Subset
- Request Status – W – Waiting to be granted, Q – Queued, O – data Online
- Size Input – Size for input data
- Size Output – Size for output data
- Date Requested – Date the request is submitted
- Date Ready – Date the request is ready for the user
- Date Purge – Date the output data is to be purged
- Compression – Compress the output data, e.g. GZ
- Request Info – Detail subsetting request constraints
- Host Name – Specify one or multiple computer nodes to process the request

Implementation of DSRQST – File Information Configuration

- File Name – An output file name created during the data processing
- Request ID – Unique ID for the request
- Source File – Link to an input file used, if available
- File Type – D for Data, O for Document, S for Software
- Status – O – data Online, R – Requested, E – Error detected
- Data Size – Output data file size
- Data Format – Output data format, e.g. netCDF
- Archive Format – Archive format, e.g. GZ
- Date – Date file is staged
- Description – Individual file description

Example – ICOADS Subset, Request Control Configuration

- Control ID – 15
- Product ID – ds540.0/41
- Control Option – A
- Command – imma_subset
- URL – /datasets/ds540.0/imma_subset.php
- Compression –
- Valid Period – 5 days
- Host Name –

Example – ICOADS Subset, Data Access

International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5, Individual Observations

ds540.0

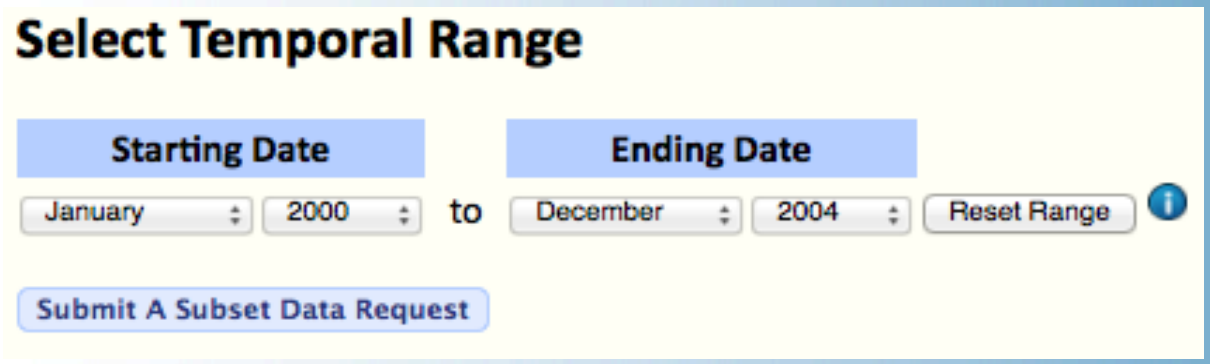
For assistance, contact Zaihua Ji (303-497-1819).

Description | Data Access | Documentation | Software

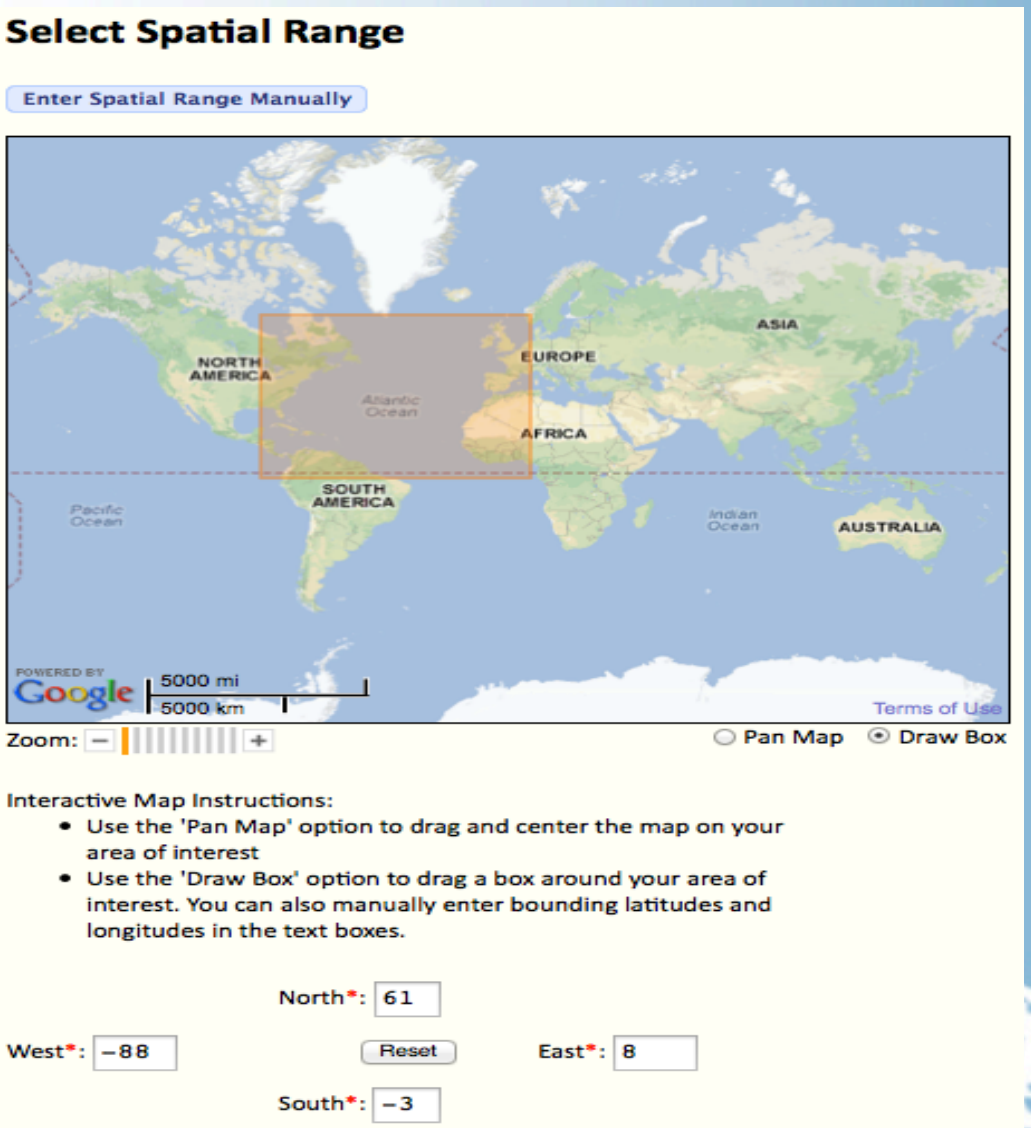
Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	NCAR-Only Access	
		Web Server Holdings	Staged Access from Tape Archive	Subsetting	Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
Union of Available Products		Web File Listing	Request Access		GLADE File Listing	HPSS File Listing
PRODUCTS	IMMA format data for ICOADS Release 2.5	Web File Listing		Get a Subset	GLADE File Listing	HPSS File Listing
	IMMA format data for ICOADS Release 2.5 Intermediate		Request Access			HPSS File Listing
	ICOADS Release 2.5.1 in experimental IMMA1 format	Web File Listing			GLADE File Listing	HPSS File Listing
	ICOADS Release 2.5.1 Intermediate in experimental IMMA1 format		Request Access			HPSS File Listing

Example – ICOADS Subset Request, Temporal Selection



Example – ICOADS Subset Request, Spatial Selection



Example – ICOADS Subset Request, Variable Selection

The IMMA Core Parameter Selection

Detailed description of the Core parameters.

Reset Core Selection

☑ YR	year UTC	☑ MO	month UTC	☑ DY	day UTC
☑ HR	hour UTC	☑ LAT	latitude	☑ LON	longitude
☐ IM	IMMA version	☐ ATTC	attm count	☐ TI	time indicator
☐ LI	latitude/long. indic.	☐ DS	ship course	☐ VS	ship speed
☐ NID	national source indic.	☐ II	ID indicator	☐ ID	identification/call sign
☐ C1	country code	☐ DI	wind direction indic.	☐ D*	wind direction
☐ WI	wind speed indicator	☐ W*	wind speed	☐ VI	VV indic.
☐ VV*	visibility	☐ WW*	present weather	☐ W1*	past weather
☐ SLP*	sea level pressure	☐ A*	characteristic of PPP	☐ PPP*	amt. pressure tend.
☐ IT	indic. for temperatures	☑ AT*	air temperature	☐ WBTI	indic. for WBT
☐ WBT*	web-bulb temperature	☐ DPTI	DPT indic.	☐ DPT*	dew-point temp.
☐ SI	SST meas. method	☑ SST*	sea surface temp.	☐ N*	total cloud amount
☐ NH*	low cloud amount	☐ CL*	low cloud type	☐ HI	H indic.
☐ H*	cloud height	☐ CM*	middle cloud type	☐ CH*	high cloud type
☐ WD*	wave direction	☐ WP*	wave period	☐ WH*	wave height
☐ SD*	swell direction	☐ SP*	swell period	☐ SH*	swell height

Submit A Subset Data Request

CISL Computational & Information Systems Laboratory

NCAR

Example – ICOADS Subset Request, Submitted

Subset Data Request 30343

Your Subset Data request has been submitted successfully. A summary of your request is given below.

Your request will be processed soon. You will be informed via email when the data is ready to be picked up.

You may check request status via link http://rda.ucar.edu/#ckrqst, for data requests you have submitted.

- If the information is **correct** no further action is need.
- If the information is **not correct**, or if you have additional comments you may email *Zaihua Ji* with corrections or comments.

```
Request Summary:
Index     : 30343
ID        : JI30343
Category  : Subset Data
Status    : Queue
Dataset   : ds540.0
Title     : International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5, Individual Observations
User      : Zaihua Ji
Email     : zji@ucar.edu
Date      : 2013-03-26
Time      : 14:25:56
Compress  : GZ
Request Detail:
Date Limits      : 200001 200412
Latitude Limits  : 3 S, 61 N
Longitude Limits : 88 W, 8 E
Filter options   : 0 1 0 0 2 1
Variable Names   : YR, MO, DY, HR, LAT, LON, AT, SST, B1
File Compression : gz
```

CISL Computational & Information Systems Laboratory

NCAR

Example – ICOADS Subset, Request Information Configuration

- Request ID – 30343
- User Email – zji@ucar.edu
- Product ID – ds540.0/41
- Request Type – S
- Request Status – Q (=> O)
- Size Input – 11825172986 (Bytes)
- Size Output – 79524274 (Bytes)
- Date Requested – 2013-03-26
- Date Ready – 2013-03-26
- Date Purge – 2013-03-31
- Request Info – dates=200001 200412&lats=3 S, 61 N&
  lons=88 W, 8 E&flts=0 1 0 0 2 1&
  vars=YR, MO, DY, HR, LAT, LON, AT, SST, B1
- Compression – GZ
- Host Name –

Example – ICOADS Subset Request, Progressing Status

Status of 1 Data Request from Zaihua Ji (zji@ucar.edu)

Index	User Email	Dataset	Type	Time Requested	Size	#File	Status
30343	zji@ucar.edu	ds540.0	Subset Data	2013-03-26 14:25:56	0.00B	63	Process - 3% built on LSF(98319/geyser03)

Example – ICOADS Subset Request, Email Notice

Zaihua Ji March 26, 2013 3:37 PM

To: Zaihua Ji

Ready for Subset Data Request '30343' of ds540.0!

Subset Data of ds540.0 - 'International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5, Individual Observations' that you have requested are ready for you to pick up. You first need to sign in to the RDA server at http://rda.ucar.edu, and then you will find your data at:

http://rda.ucar.edu/#dsrqst/JI30343/

Request Detail:
Date Limits : 200001 200412
Latitude Limits : 3 S, 61 N
Longitude Limits : 88 W, 8 E
Filter options : 0 1 0 0 2 1
Variable Names : YR, MO, DY, HR, LAT, LON, AT, SST, B1
File Compression : gz

Your data will remain on our system for 5 days. If this is not sufficient time for you to retrieve your data, please let me know as soon as possible, so that I can prevent the data files from being purged too soon.

If you have any questions related to this data request, please let me know by replying to this email.

Sincerely,

Zaihua Ji
NCAR/CISL/Data Support Section
Phone: (303)-497-1819
Email: zji@ucar.edu
Web: http://rda.ucar.edu

Example – ICOADS Subset, File Information

- File Name – ICOADS.200001.200412_1.gz
- Request ID – 30343
- Source File –
- File Type – D
- Status – O
- Data Size – 5920104
- Data Format – ASCII
- Archive Format – GZ
- Date – 2013-03-23
- Description -

Example – ICOADS Subset Request, Data Web Page

Subset Data Requested From DS540.0

Go to DS540.0 Home Page.

The Subset Data requested by Zaihua Ji (zji@ucar.edu) from ds540.0 - 'International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5, Individual Observations' are listed below. The data files will remain available online for 5 days (until 2013-03-31 15:35:49).If you need additional time to complete the download, please notify *Zaihua Ji*. After you have completed the download, please also let me know by clicking this button

I Have Finished Download, Please Purge the Request

```
Date Limits      : 200001 200412
Latitude Limits  : 3 S, 61 N
Longitude Limits : 88 W, 8 E
Filter options   : 0 1 0 0 2 1
Variable Names   : YR, MO, DY, HR, LAT, LON, AT, SST, B1
File Compression : gz
```

Show Selected Files/Get As a Tar File | Perl Download Script | Csh Download Script

- Total **16 Files (79.52M)** are listed below
- Click a file name to download a single file
- Select one or multiple files to get a download script
- Select multiple data files to download as a single tar file
- Currently **0 File** selected Clear Selection

INDEX	File Name	Size	Type	Data Format	Archive Format	Date Online
1	ICOADS.200001.200412_1.gz	5.92M	Data	ASCII	GZ	03/26/2013
2	ICOADS.200001.200412_2.gz	6.65M	Data	ASCII	GZ	03/26/2013
3	ICOADS.200001.200412_3.gz	5.94M	Data	ASCII	GZ	03/26/2013
4	ICOADS.200001.200412_4.gz	6.83M	Data	ASCII	GZ	03/26/2013
5	ICOADS.200001.200412_5.gz	6.08M	Data	ASCII	GZ	03/26/2013
6	ICOADS.200001.200412_6.gz	5.82M	Data	ASCII	GZ	03/26/2013
7	ICOADS.200001.200412_7.gz	5.59M	Data	ASCII	GZ	03/26/2013
8	ICOADS.200001.200412_8.gz	5.96M	Data	ASCII	GZ	03/26/2013
9	ICOADS.200001.200412_9.gz	5.35M	Data	ASCII	GZ	03/26/2013
10	ICOADS.200001.200412_10.gz	6.04M	Data	ASCII	GZ	03/26/2013
11	ICOADS.200001.200412_11.gz	6.19M	Data	ASCII	GZ	03/26/2013
12	ICOADS.200001.200412_12.gz	5.84M	Data	ASCII	GZ	03/26/2013
13	ICOADS.200001.200412_13.gz	6.00M	Data	ASCII	GZ	03/26/2013
14	readme_imma.ji30343	5.08K	Document	TEXT		03/26/2013
15	R2.5-imma.pdf	768.49K	Document	PDF		03/26/2013
16	R2.5-imma_short.pdf	549.26K	Document	PDF		03/26/2013

Show Selected Files/Get As a Tar File | Perl Download Script | Csh Download Script

CISL Computational & Information Systems Laboratory

NCAR

Metrics for Data Services, March 2012 - March 2013				
Processed by Request Controller (293 Products in 79 Datasets)			Direct File Downloads (1100 Products in 650 Datasets)	
#Users	Served(TB)	Processed(TB)	#users	Served(TB)
3464	274	9594	7402	660

Conclusion

- A scalable and stable data request controller, DSRQST, is implemented in three levels of programming configuration (recorded in RDADB)
- Supports multiple request types, including data subsetting, format converting, and data staging onto disk-based systems that are convenient for access
- Easy links to customized web interface to accept users requests and background command to process the data
- Controller runs on distributed servers for queued requests
- Failed request processes are detected and reprocessed
- Identical data file is shared by multiple requests
- Purge temporarily staged data and record usage metrics automatically

Future Works

- Expand to include more services
- Intergrade with OPeNDAP
- And others?

Zaihua Ji
NCAR/CISL/Data Support Section
Phone: (303)-497-1819
Email: zji@ucar.edu
Web: http://rda.ucar.edu