Data Publication Implementation and Metrics

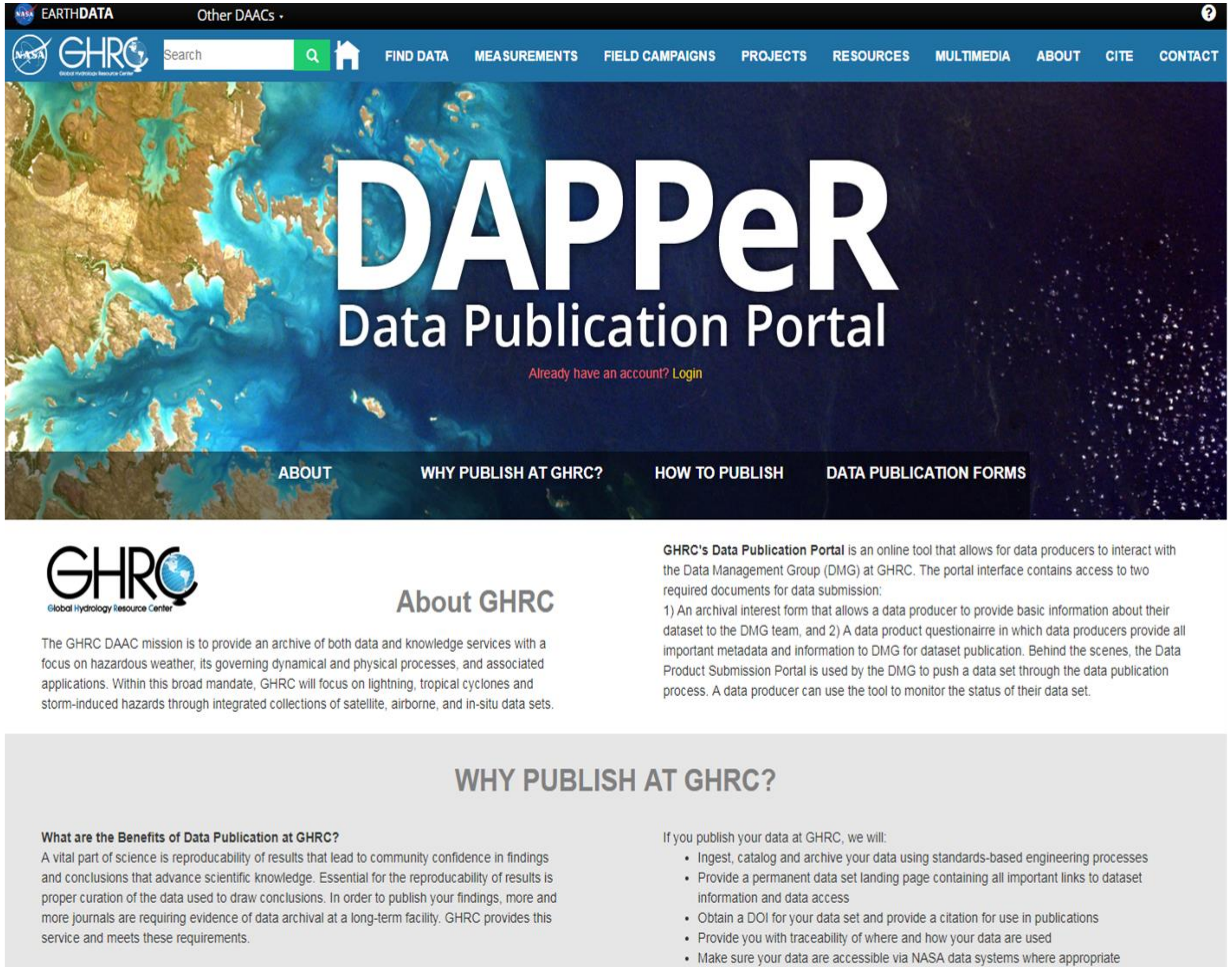
Deborah Smith

2017 GHRC User Working Group Meeting
Sept 26-27, 2017

itsc
INFORMATION TECHNOLOGY AND SYSTEMS CENTER

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

What is DAPPeR?



DAPPeR

Data Publication Portal

Already have an account? Login

About GHRC

The GHRC DAAC mission is to provide an archive of both data and knowledge services with a focus on hazardous weather, its governing dynamical and physical processes, and associated applications. Within this broad mandate, GHRC will focus on lightning, tropical cyclones and storm-induced hazards through integrated collections of satellite, airborne, and in-situ data sets.

GHRC's Data Publication Portal is an online tool that allows for data producers to interact with the Data Management Group (DMG) at GHRC. The portal interface contains access to two required documents for data submission:
1) An archival interest form that allows a data producer to provide basic information about their dataset to the DMG team, and 2) A data product questionairre in which data producers provide all important metadata and information to DMG for dataset publication. Behind the scenes, the Data Product Submission Portal is used by the DMG to push a data set through the data publication process. A data producer can use the tool to monitor the status of their data set.

WHY PUBLISH AT GHRC?

What are the Benefits of Data Publication at GHRC?
A vital part of science is reproducability of results that lead to community confidence in findings and conclusions that advance scientific knowledge. Essential for the reproducability of results is proper curation of the data used to draw conclusions. In order to publish your findings, more and more journals are requiring evidence of data archival at a long-term facility. GHRC provides this service and meets these requirements.

If you publish your data at GHRC, we will:

- Ingest, catalog and archive your data using standards-based engineering processes
- Provide a permanent data set landing page containing all important links to dataset information and data access
- Obtain a DOI for your data set and provide a citation for use in publications
- Provide you with traceability of where and how your data are used
- Make sure your data are accessible via NASA data systems where appropriate

DAPPeR is an online publication tool based on ORNL SAuS

- Put into operation in May 2017

DAPPeR web page contains description of publication process and a direct link to the data archival interest form

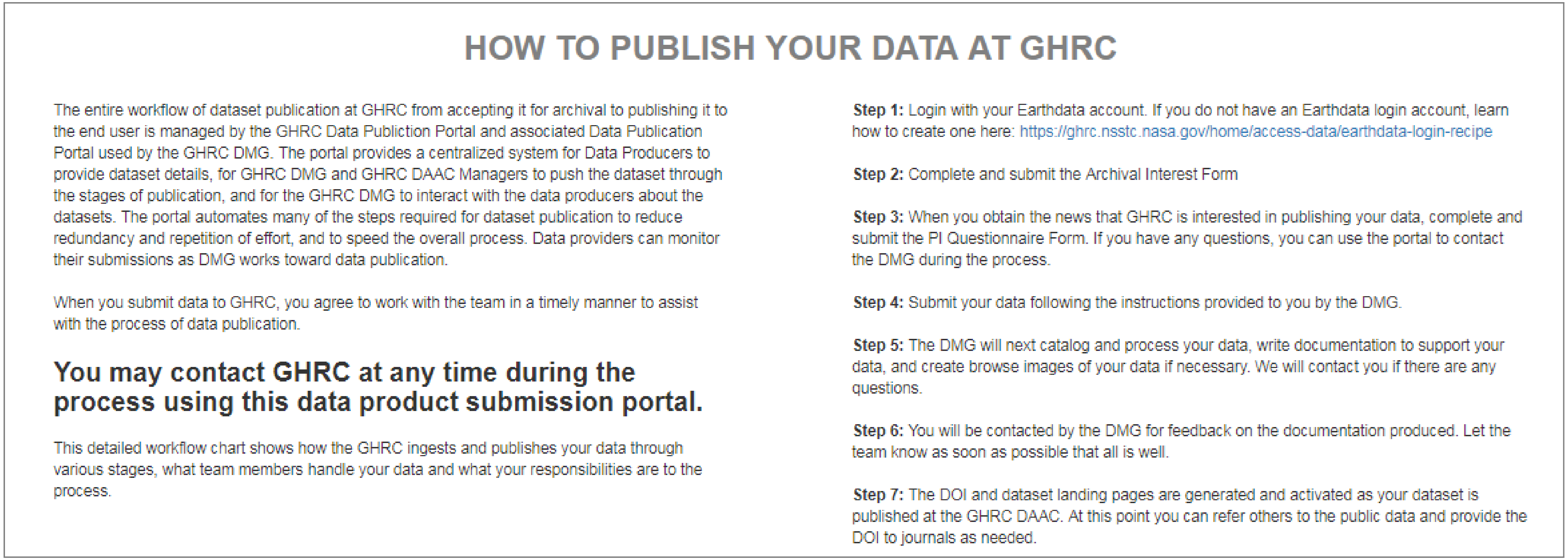
- Made public in August 2017

The secure portal interface has access for both data providers and GHRC personnel

DAPPeR URL: https://ghrc.nsstc.nasa.gov/data-publication/

DAPPeR: Dataset Submission on Web

- Description of publication process is public with needed links for more information and form access
- Feedback link is available for questions/comments



HOW TO PUBLISH YOUR DATA AT GHRC

The entire workflow of dataset publication at GHRC from accepting it for archival to publishing it to the end user is managed by the GHRC Data Publiction Portal and associated Data Publication Portal used by the GHRC DMG. The portal provides a centralized system for Data Producers to provide dataset details, for GHRC DMG and GHRC DAAC Managers to push the dataset through the stages of publication, and for the GHRC DMG to interact with the data producers about the datasets. The portal automates many of the steps required for dataset publication to reduce redundancy and repetition of effort, and to speed the overall process. Data providers can monitor their submissions as DMG works toward data publication.

When you submit data to GHRC, you agree to work with the team in a timely manner to assist with the process of data publication.

You may contact GHRC at any time during the process using this data product submission portal.

This detailed workflow chart shows how the GHRC ingests and publishes your data through various stages, what team members handle your data and what your responsibilities are to the process.

Step 1: Login with your Earthdata account. If you do not have an Earthdata login account, learn how to create one here: https://ghrc.nsstc.nasa.gov/home/access-data/earthdata-login-recipe

Step 2: Complete and submit the Archival Interest Form

Step 3: When you obtain the news that GHRC is interested in publishing your data, complete and submit the PI Questionnaire Form. If you have any questions, you can use the portal to contact the DMG during the process.

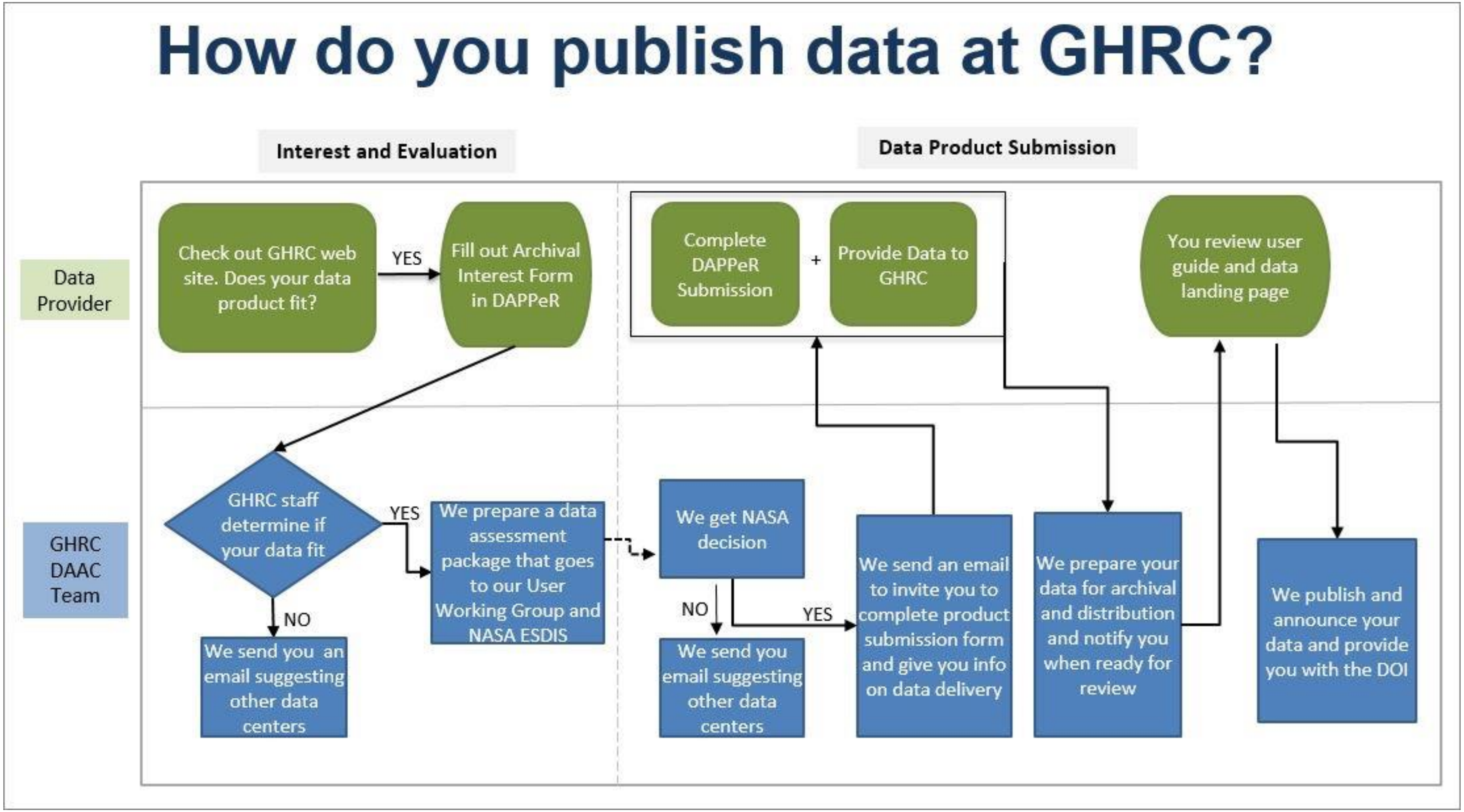
Step 4: Submit your data following the instructions provided to you by the DMG.

Step 5: The DMG will next catalog and process your data, write documentation to support your data, and create browse images of your data if necessary. We will contact you if there are any questions.

Step 6: You will be contacted by the DMG for feedback on the documentation produced. Let the team know as soon as possible that all is well.

Step 7: The DOI and dataset landing pages are generated and activated as your dataset is published at the GHRC DAAC. At this point you can refer others to the public data and provide the DOI to journals as needed.

Swimlane Chart for Data Provider



How do you publish data at GHRC?

Interest and Evaluation

Data Product Submission

Data Provider

GHRC DAAC Team

Check out GHRC web site. Does your data product fit?

YES

Fill out Archival Interest Form in DAPPeR

GHRC staff determine if your data fit

NO

We send you an email suggesting other data centers

YES

We prepare a data assessment package that goes to our User Working Group and NASA ESDIS

We get NASA decision

NO

We send you email suggesting other data centers

YES

We send an email to invite you to complete product submission form and give you info on data delivery

Complete DAPPeR Submission

+

Provide Data to GHRC

We prepare your data for archival and distribution and notify you when ready for review

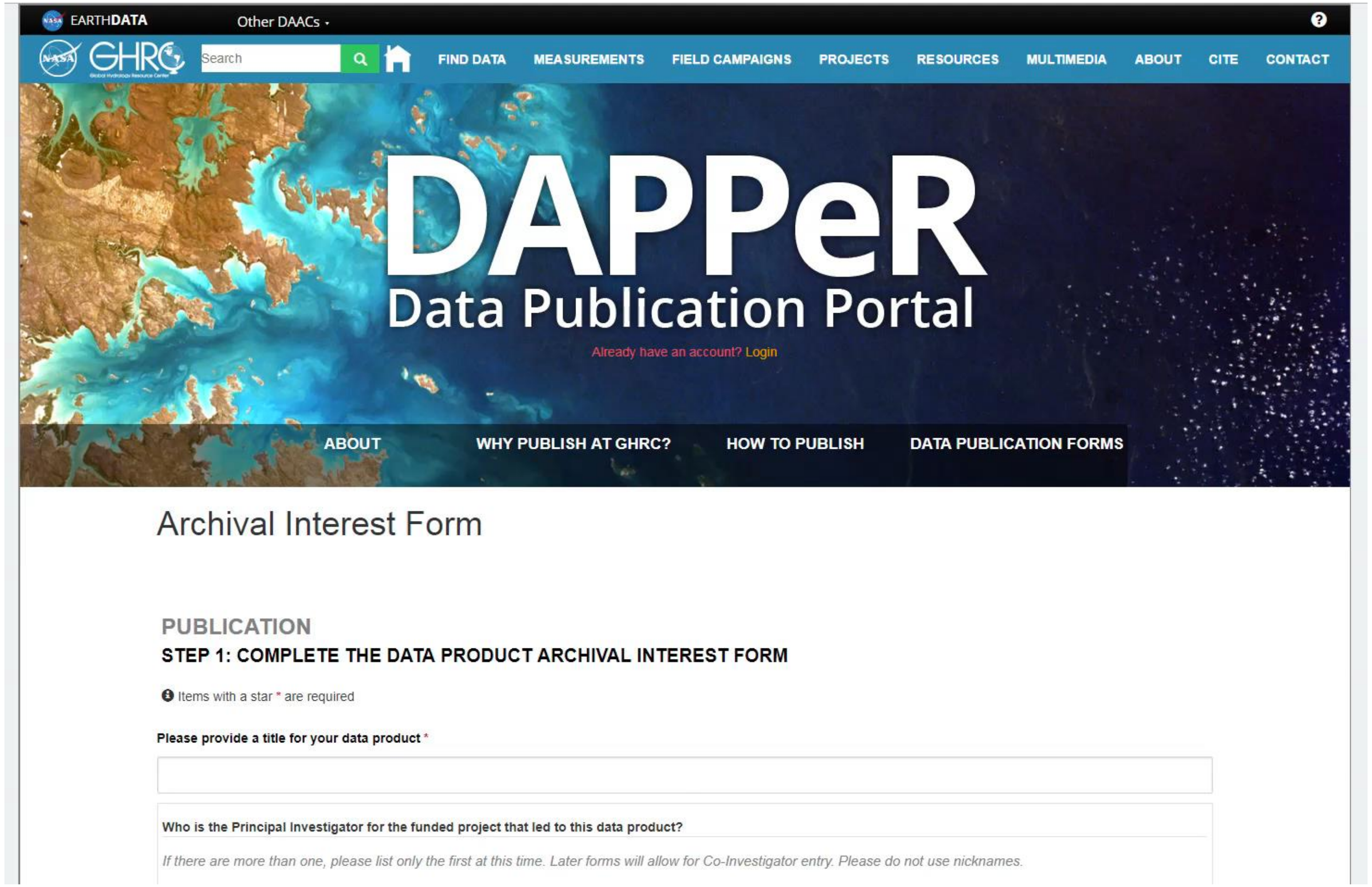
You review user guide and data landing page

We publish and announce your data and provide you with the DOI

Data Publication Forms

- Archival Interest Form
  - Used to collect information about and relevancy of a proposed dataset to GHRC
  - https://ghrc.nsstc.nasa.gov/data-publication/node/add/archival_interest_form

- Data Product Questionnaire
  - Secure access for data providers to submit detailed dataset information once the dataset has cleared the review process
  - Data producers are required to complete most questions in order for dataset to be accepted



Already have an account? Login

Archival Interest Form

PUBLICATION

STEP 1: COMPLETE THE DATA PRODUCT ARCHIVAL INTEREST FORM

Items with a star * are required

Please provide a title for your data product *

Who is the Principal Investigator for the funded project that led to this data product?

If there are more than one, please list only the first at this time. Later forms will allow for Co-Investigator entry. Please do not use nicknames.

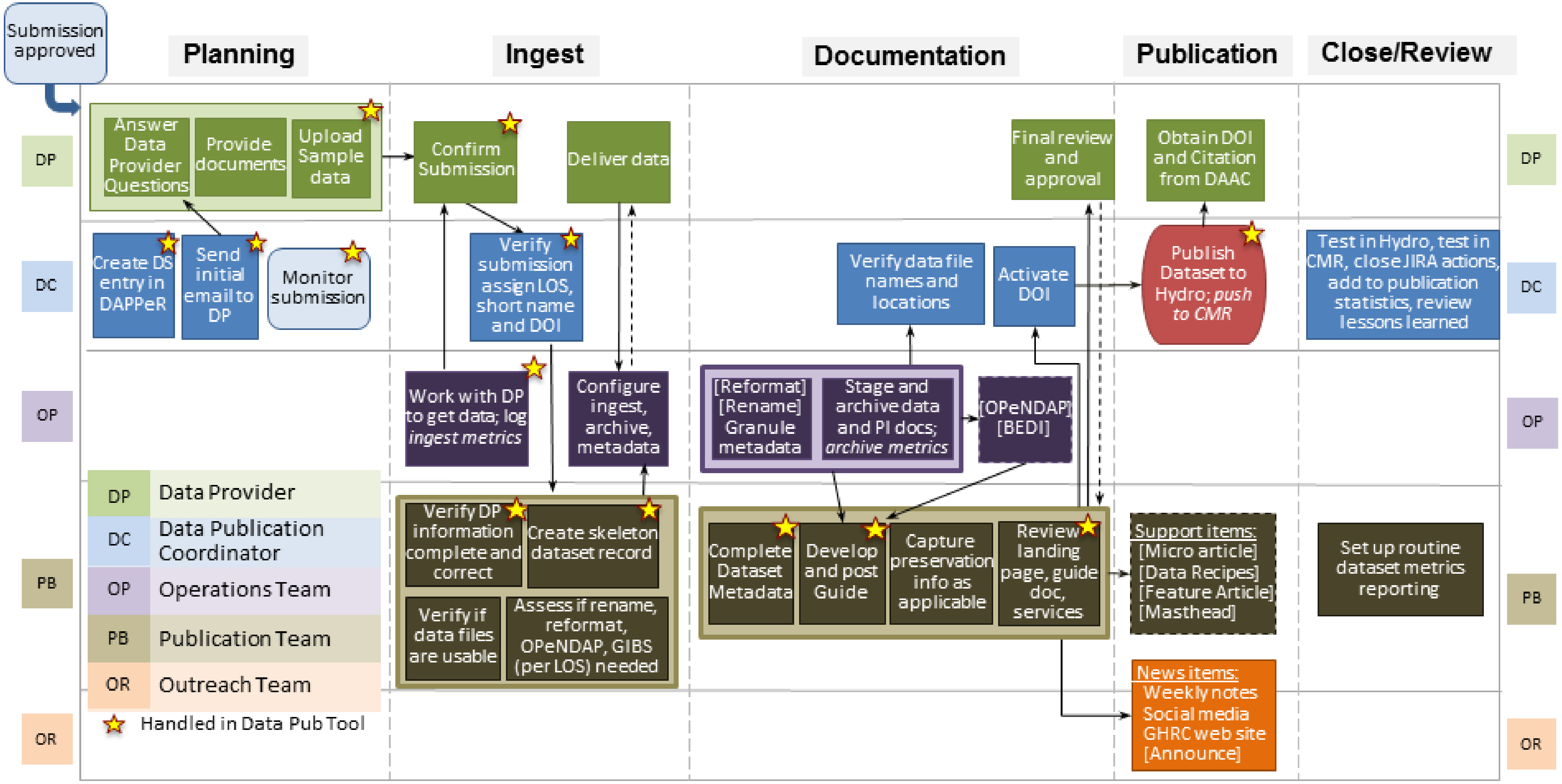
GHRC Data Publication Changes

The GHRC data publication process underwent great change in 2017

- A new database and publication tool DAPPeR was installed, tested, and implemented during the beginning of 2017
- The publication team had to be trained in DAPPeR, including clear understanding of metadata definitions so that consistent information would be entered into the database by everyone
- During Jan - Apr, the publication team pre-prepared datasets
- Operational use of DAPPeR started in May 2017
- In July, Marilyn retired, resulting in a change of leadership
- In 4.5 months, we can already see an increase in publication rate from the new tools and processes

Data Publication Swimlane Chart Updated

GHRC Data Publication Swimlanes



Submission approved

Planning

Ingest

Documentation

Publication

Close/Review

DP: Answer Data Provider Questions; Provide documents; Upload Sample data; Confirm Submission; Deliver data; Final review and approval; Obtain DOI and Citation from DAAC

DC: Create DS entry in DAPPeR; Send initial email to DP; Monitor submission; Verify submission assign LOS, short name and DOI; Verify data file names and locations; Activate DOI; Publish Dataset to Hydro; *push to CMR*; Test in Hydro, test in CMR, close JIRA actions, add to publication statistics, review lessons learned

OP: Work with DP to get data; log *ingest metrics*; Configure ingest, archive, metadata; [Reformat] [Rename] Granule metadata; Stage and archive data and PI docs; *archive metrics*; [OPeNDAP] [BEDI]

PB: Verify DP information complete and correct; Create skeleton dataset record; Verify if data files are usable; Assess if rename, reformat, OPeNDAP, GIBS (per LOS) needed; Complete Dataset Metadata; Develop and post Guide; Capture preservation info as applicable; Review landing page, guide doc, services; Support items: [Micro article] [Data Recipes] [Feature Article] [Masthead]; Set up routine dataset metrics reporting

OR: News items: Weekly notes Social media GHRC web site [Announce]

DP	Data Provider
DC	Data Publication Coordinator
OP	Operations Team
PB	Publication Team
OR	Outreach Team

☆ Handled in Data Pub Tool

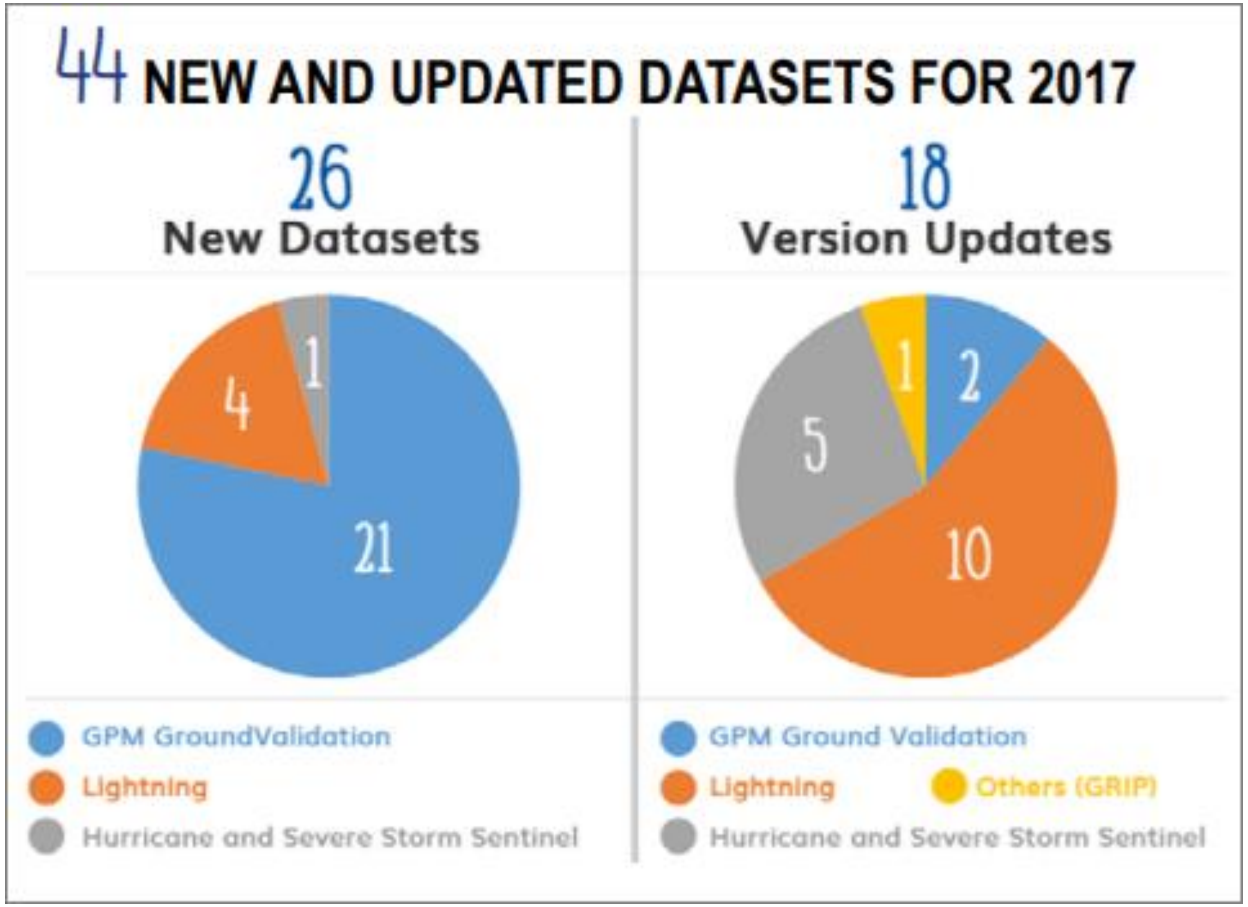
* adapted from ORNL DAAC's Swimlanes

GHRC Data Publications for 2017

Progress has been made on publishing datasets already delivered to GHRC

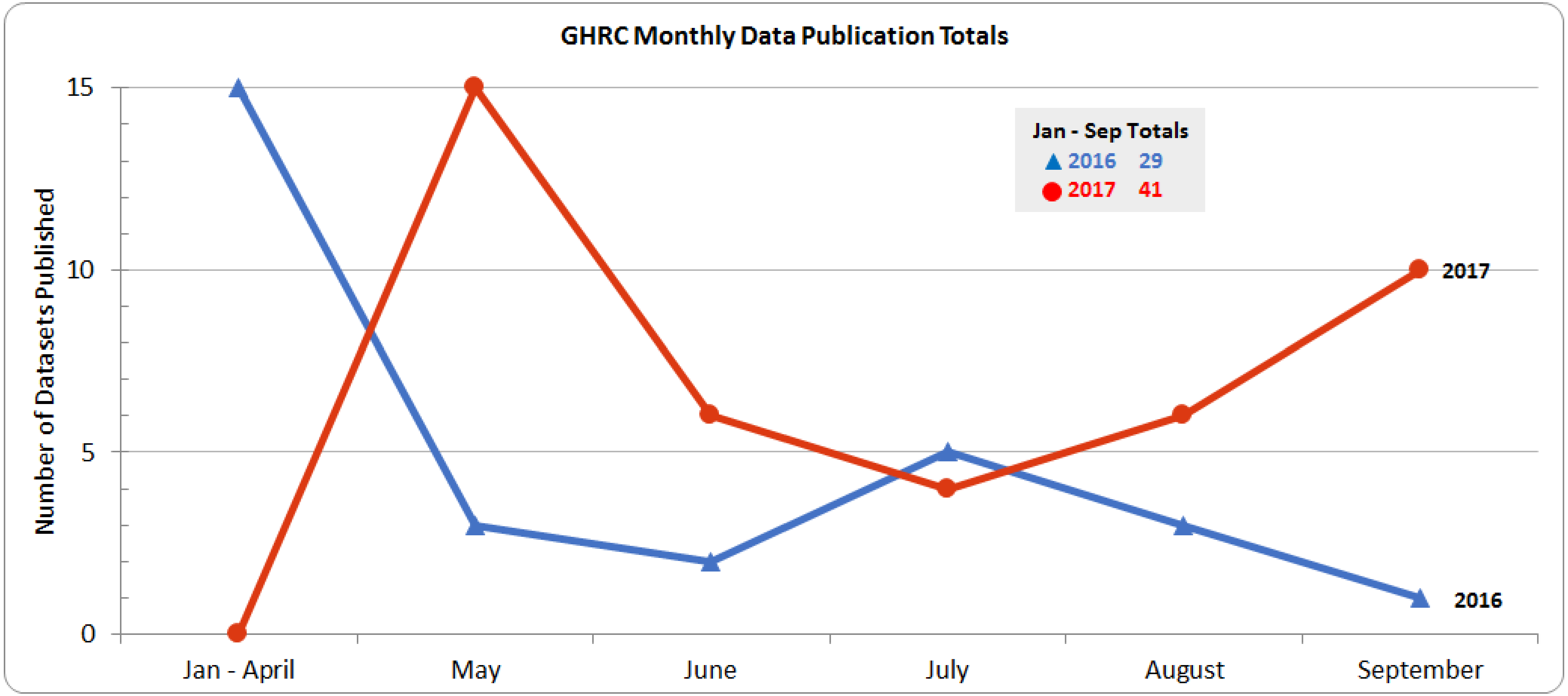
Focus has been on first publishing the highest level of service datasets

- ISS LIS NRT and Non-QC datasets published
- OLYMPEX GPM Ground Validation
  - 50% of OLYMPEX campaign published
- All of HS3 core datasets are now published



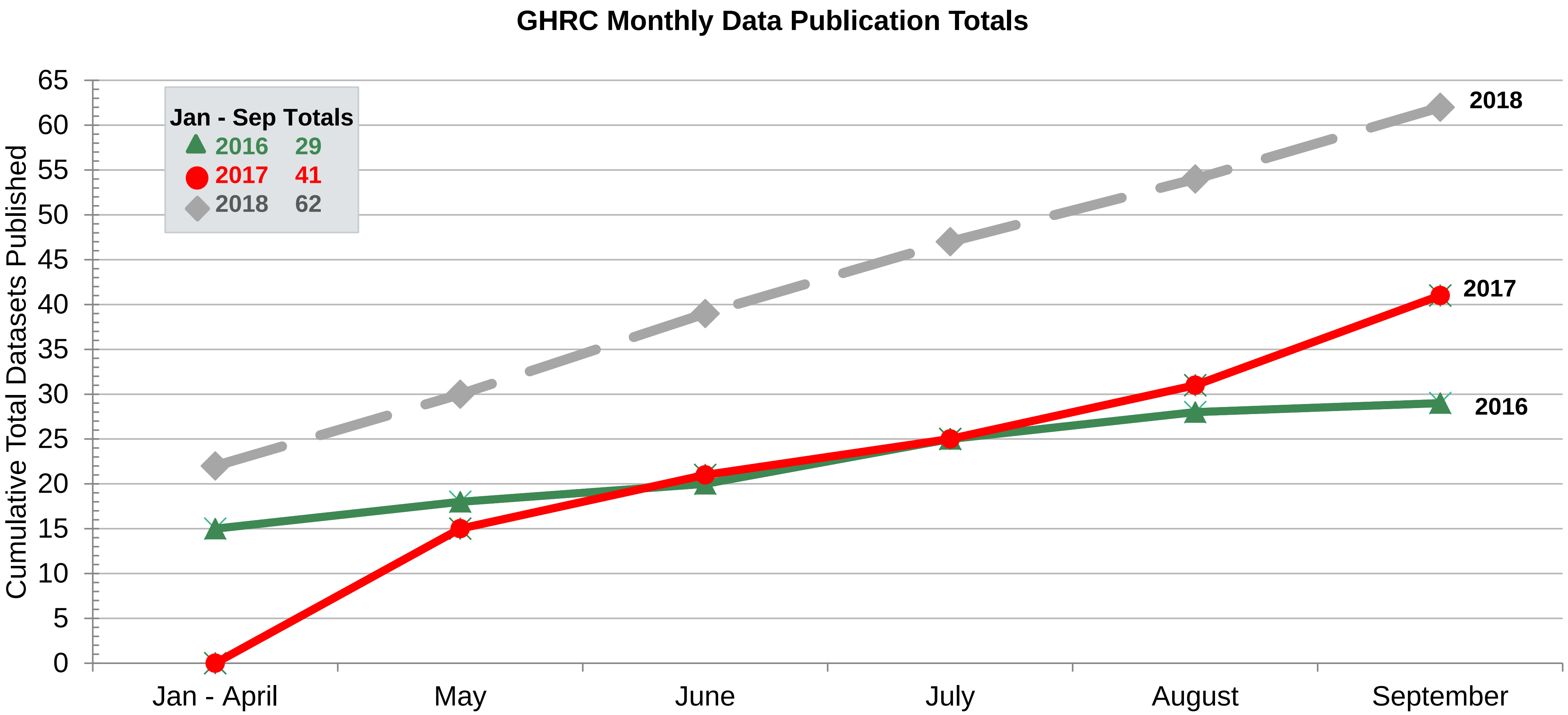
Monthly Data Publication Rates Already Improved

Comparison of dataset publication monthly totals for Jan - Sep 2016 and 2017



Monthly Data Publication Rates Already Improved

Comparison of dataset publication cumulative totals per year, 2016 -2018



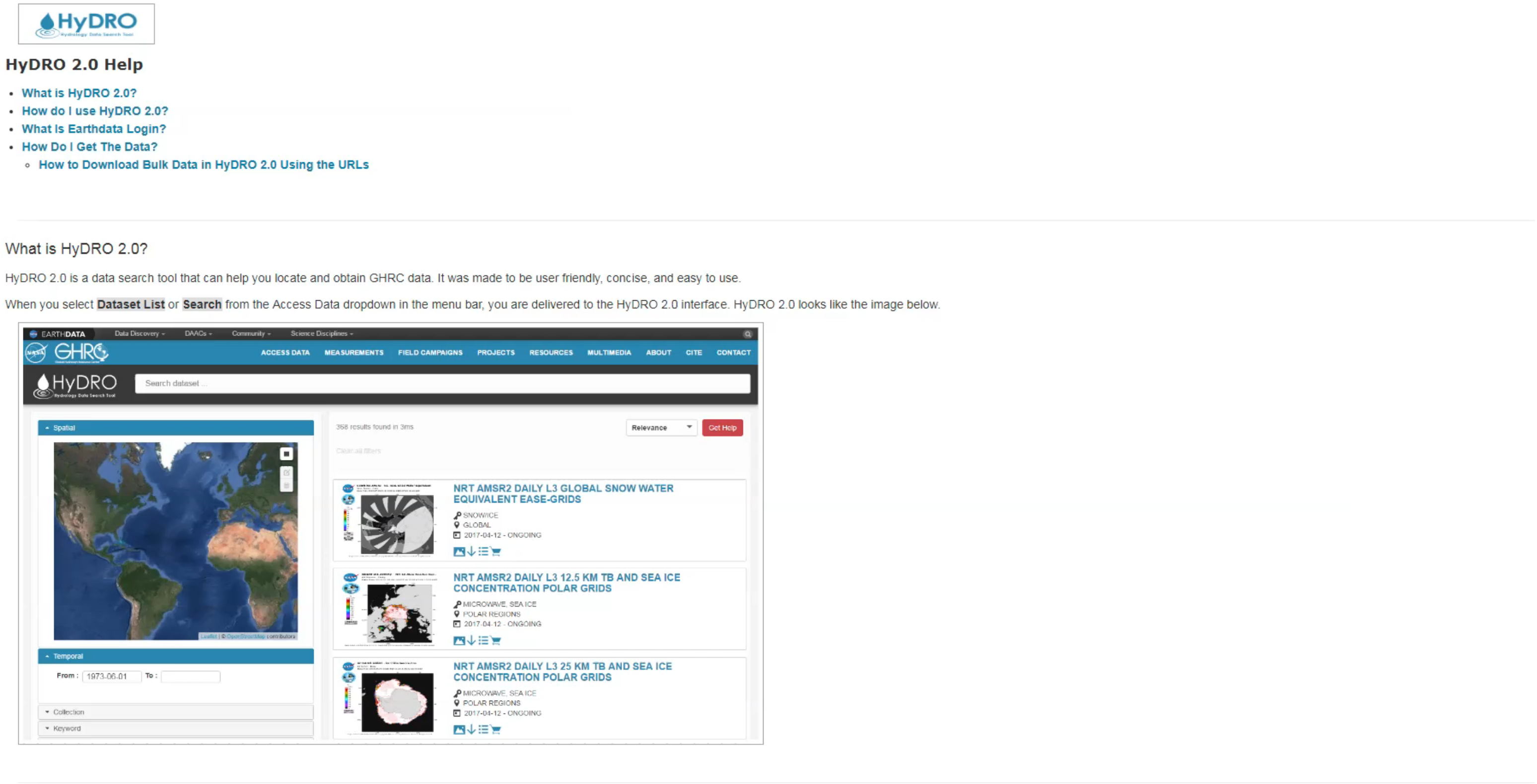
Publication Statistics From DAPPeR

- By using DAPPeR we will be able to record effort and time required for each dataset publication

- We anticipate to have statistics such as the following for next year's report:
    - avg time from start of effort to publication
    - avg time from ingest of dataset to publication
    - avg time for PI questionnaire completion
    - avg time per each step of publication
    - analyses by year, quarter, month
    - summary info on email communications, such as how many datasets required extra contact for publication

HyDRO 2.0

- At last UWG meeting, we demonstrated our new HyDRO 2.0
  - Tool used to locate and access data from GHRC archive
  - New features made it easier for users to narrow down searches and find what they want
- HyDRO 2.0 was made fully operational in 2017
  - •A YouTube instructional Video is provided to instruct on how to use HyDRO 2.0
- We incorporated a 'Get Help' button
- One single issue remains - Users having trouble downloading multiple files of data



HyDRO 2.0 Help

- What is HyDRO 2.0?
- How do I use HyDRO 2.0?
- What Is Earthdata Login?
- How Do I Get The Data?
  - How to Download Bulk Data in HyDRO 2.0 Using the URLs

What is HyDRO 2.0?

HyDRO 2.0 is a data search tool that can help you locate and obtain GHRC data. It was made to be user friendly, concise, and easy to use.

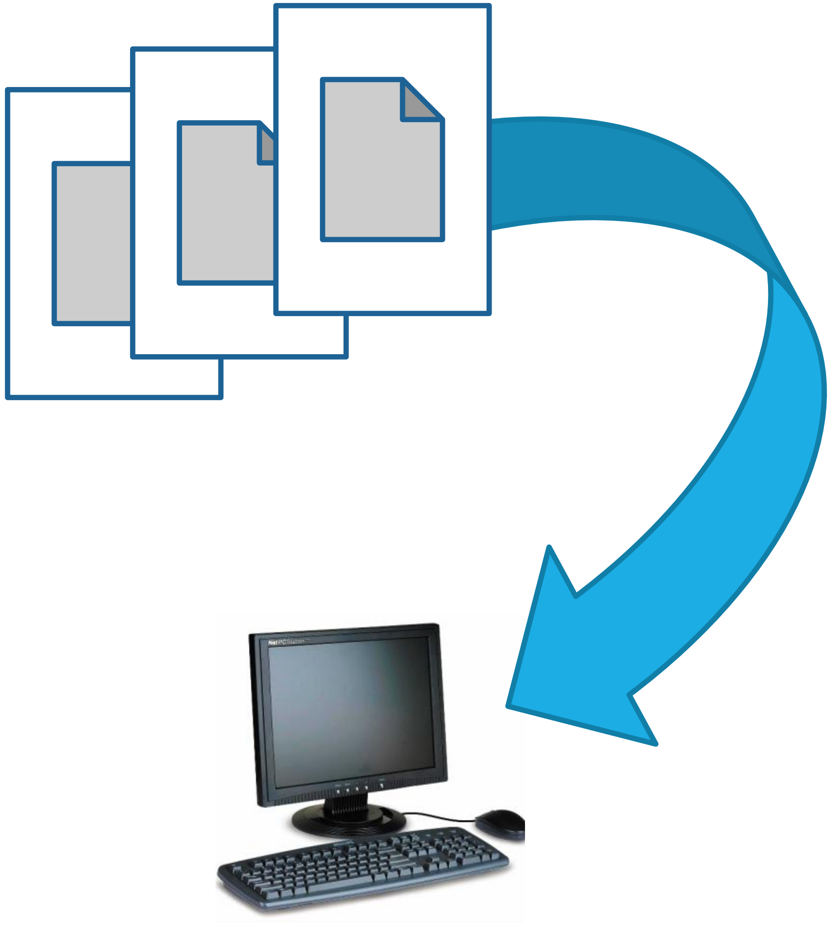
When you select **Dataset List** or **Search** from the Access Data dropdown in the menu bar, you are delivered to the HyDRO 2.0 interface. HyDRO 2.0 looks like the image below.

NRT AMSR2 DAILY L3 GLOBAL SNOW WATER EQUIVALENT EASE-GRIDS
SNOW/ICE
GLOBAL
2017-04-12 - ONGOING

NRT AMSR2 DAILY L3 12.5 KM TB AND SEA ICE CONCENTRATION POLAR GRIDS
MICROWAVE, SEA ICE
POLAR REGIONS
2017-04-12 - ONGOING

NRT AMSR2 DAILY L3 25 KM TB AND SEA ICE CONCENTRATION POLAR GRIDS
MICROWAVE, SEA ICE
POLAR REGIONS
2017-04-12 - ONGOING

Bulk Data Download Issues at GHRC



- Current process: Utilize WGET with a provided URL file to bulk download data via HTTPS
- A text file of URLs is downloaded by the user thru HyDRO 2.0, then user runs WGET calling the script
  - positives: works with Windows and Linux/Unix
  - negatives: requires user to install WGET
- WGET asks the user for Earthdata credentials before data download
- We are currently investigating other ways to bulk download data
- We have looked into Earthdata Drive, but had complications
- We have examined what other NASA DAAC's have for their own bulk data downloading tool. Various approaches exist
- We are interested in finding a way to converge all of the diverse NASA DAAC bulk data download to create one standard bulk data download tool that can be used with Earthdata login across all DAACs

Questions?

2017 GHRC User Working Group Meeting
Sept 26-27, 2017

NASA

itsc
INFORMATION TECHNOLOGY
AND SYSTEMS CENTER

UAH
THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE