Strategies for Curating Big Data in the Heterogeneous Long Tail:

Examples from the USGS ScienceBase Repository and SEAD Data Services

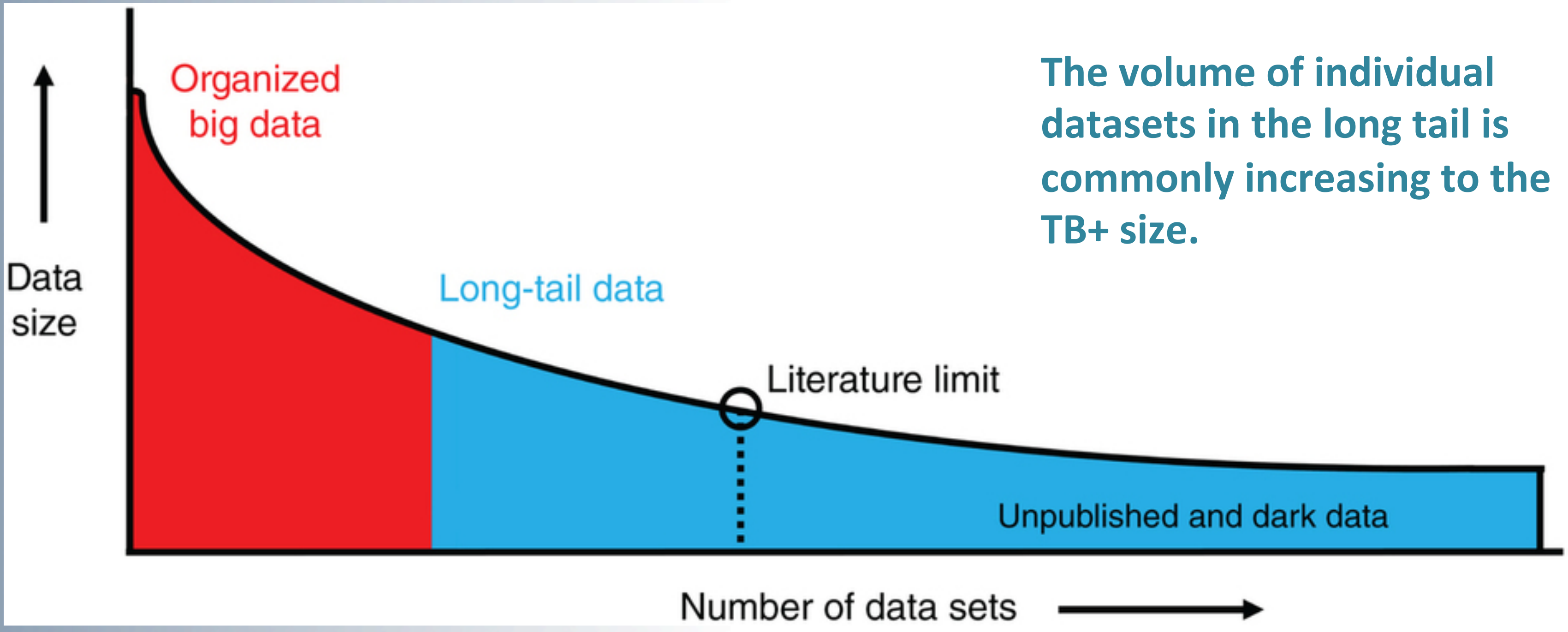
Leslie Hsu[1], Drew Ignizio[1], and James Myers[2]

[1]U.S. Geological Survey, [2]University of Michigan

Geological Society of America Annual Meeting 2016

What do we mean by

Big Data in the Heterogeneous Long Tail



The volume of individual datasets in the long tail is commonly increasing to the TB+ size.

Ferguson et al., 2014

What do we mean by

Curating

- *organization and integration*
- *annotation*
- *publication and presentation*
- *value of the data is maintained*
- ***available for** reuse and preservation*

[- Wikipedia]

Curating big data in the long tail - why now?

New Journal Review Criteria

We ask that reviewers do the following to ensure compliance with AGU's Data Policy, which requires authors to include information on data availability regarding the paper.

Read each Acknowledgments section carefully to verify that ALL data used in the research have been included

Check any hyperlinks that have been provided in the Acknowledgments to verify the accessibility of data

Report any failure to comply with the data policy when submitting a review or making a recommendation to the editor

[AGU journal review criteria]

Curating big data in the long tail - why now?

New Federal Policies for Public Access to Data

...beginning Oct. 1, 2016, the USGS will require digital research data collected with USGS funds meet the following requirements:

Scientific data that are used to support the conclusions in scholarly publications will be made available free-of-charge for public access simultaneously **with or prior to the release** of an associated scholarly publication...

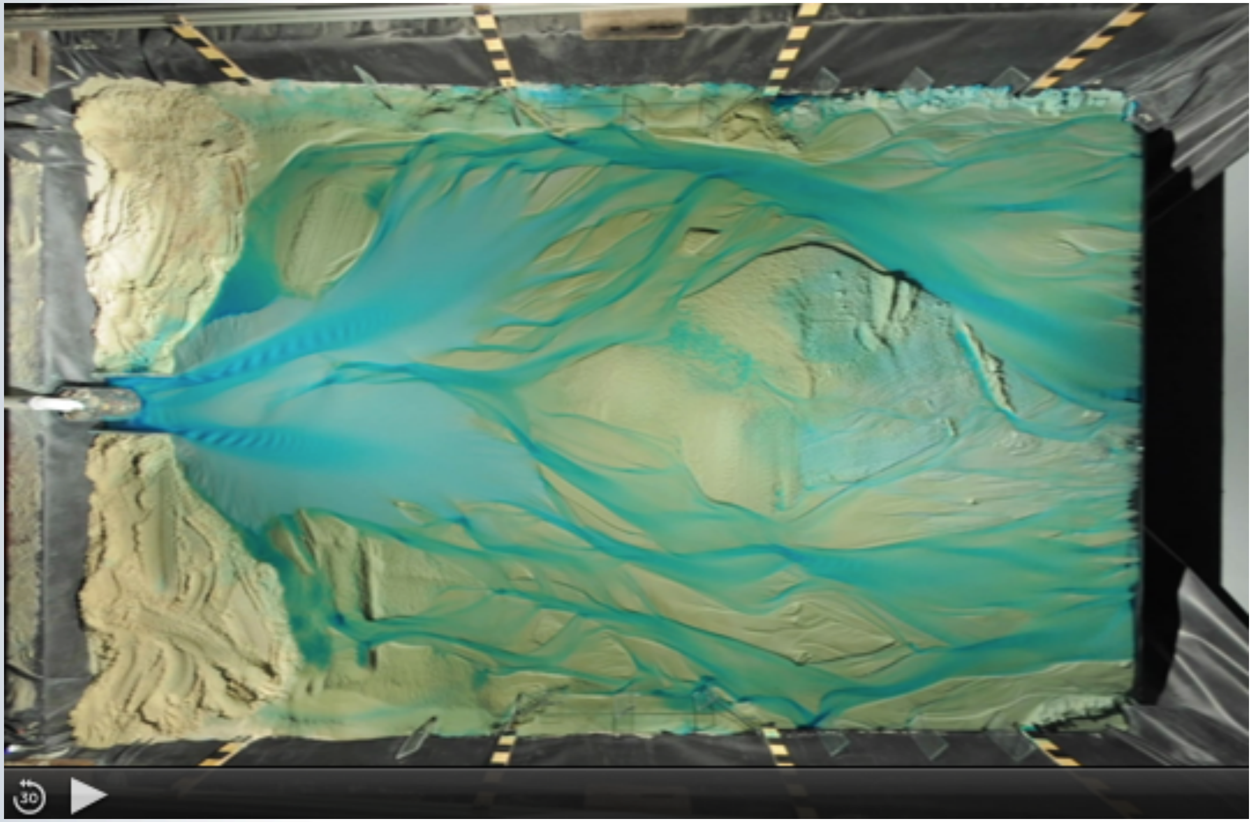
[Public Access to Results of Federally Funded Research at the U.S. Geological Survey]

How big is the data?

Bufe et al., 2016, Fluvial bevelling of topography controlled by lateral channel mobility and uplift rate. *Nature Geoscience*

Total Size:
30.96 GB
Number of Files:
11883
Largest File:
870.56 MB

My … experiments are about to be published so I just wondered about data repositories –
Last I heard [you] cannot accommodate 20 GB of images and files - is that still correct?



How big is the data?

Collins and Jibeson, 2015, Assessment of Existing and Potential Landslide Hazards Resulting from the April 25, 2015 Gorkha, Nepal Earthquake Sequence, *USGS Open File Report*

Total Size:
109 GB

We collected approximately 6,000 still-photo images of landslide-affected regions and video coverage of approximately 1,000 km of flight path…

How big is the data?

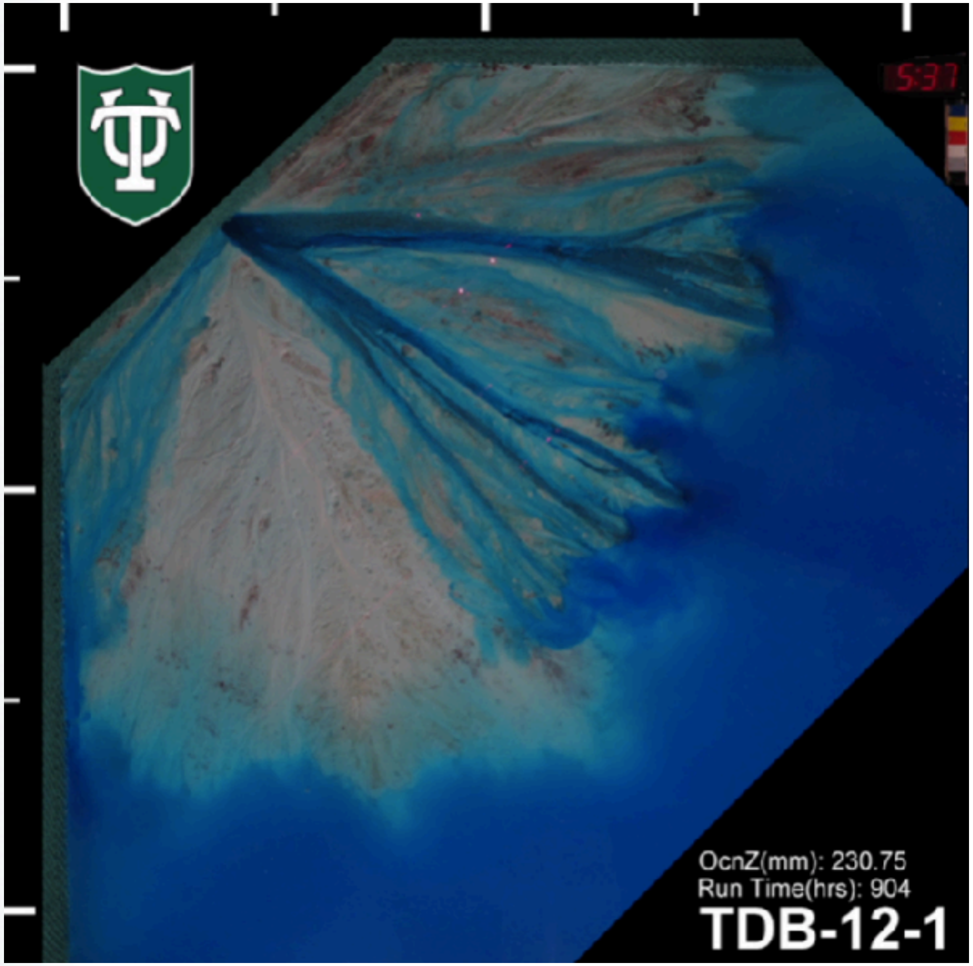
Tulane Sediment Dynamics Group

Total Size: 842 GB

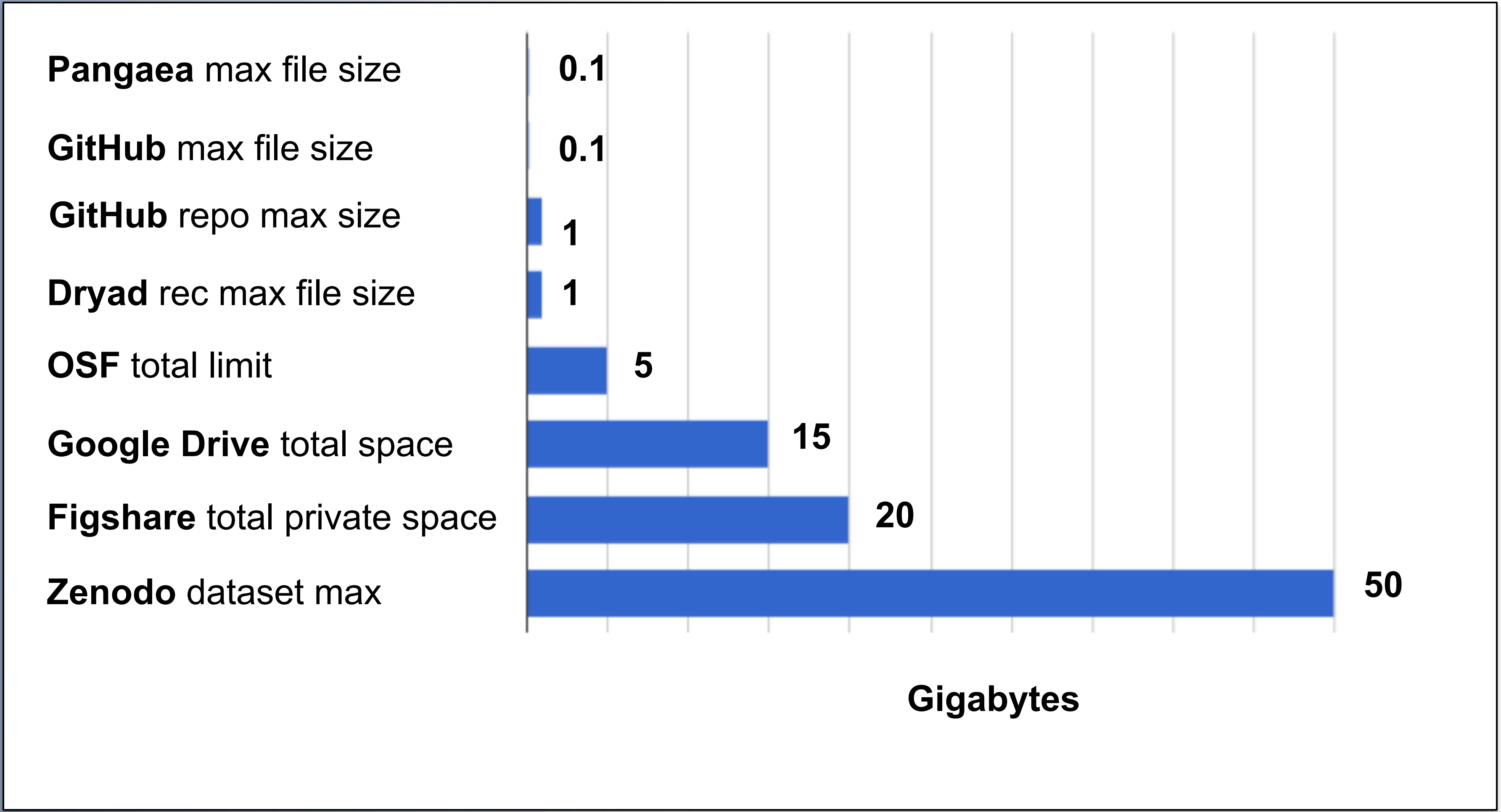
Number of Files: 3312

Largest File: 4 GB

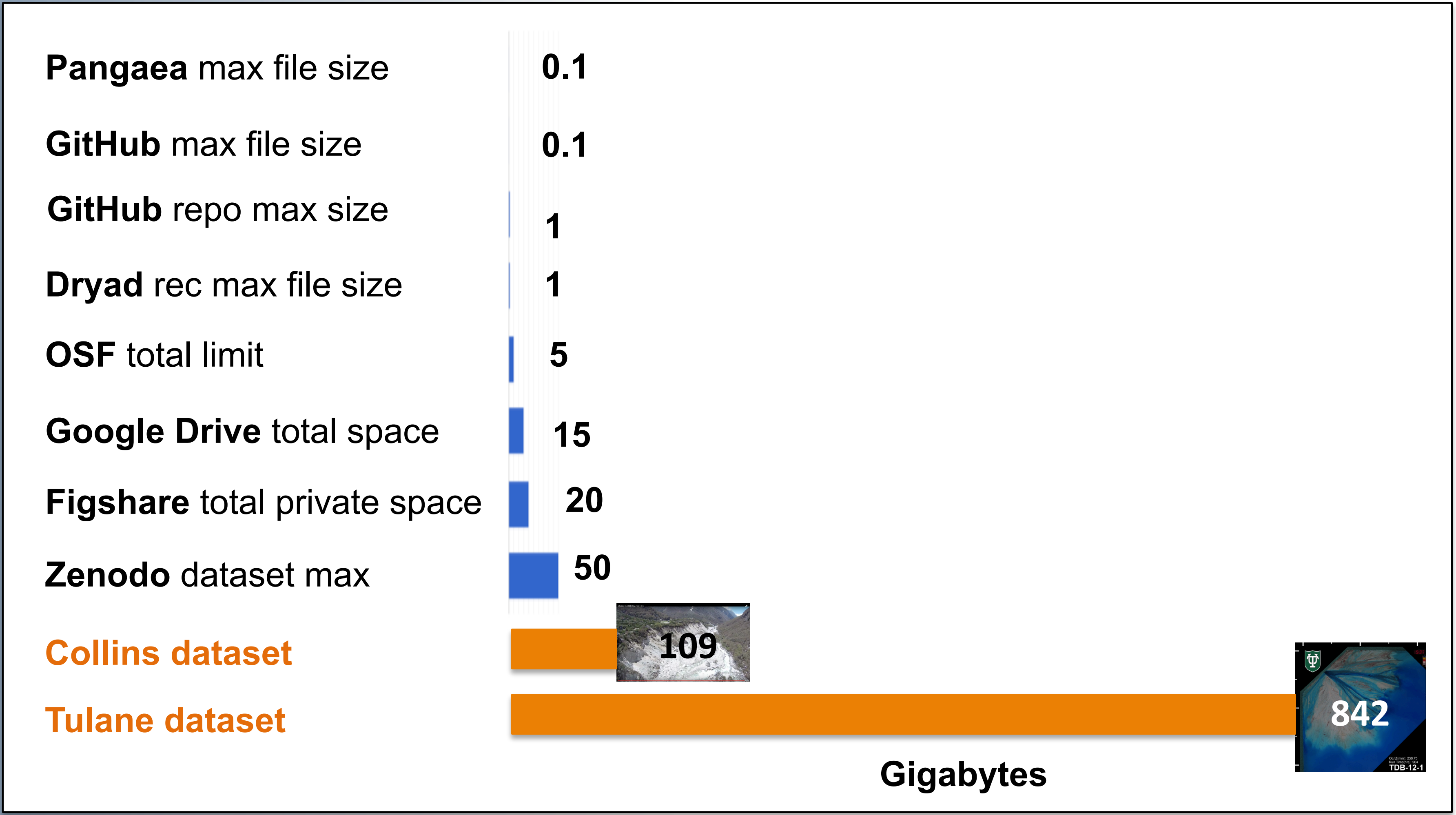
The total size of my experimental data is about 6TB.



Storage capability of some common repositories (as of Sept 25, 2016)



Storage capability of some common repositories (as of Sept 25, 2016)



Communities of Practice help find solutions

- Collection and prioritization of data curation needs
- Two-way communication between data system developers and users
- Develop disciplinary "flavor” of solutions

SEAD Data Services, sead-data.net

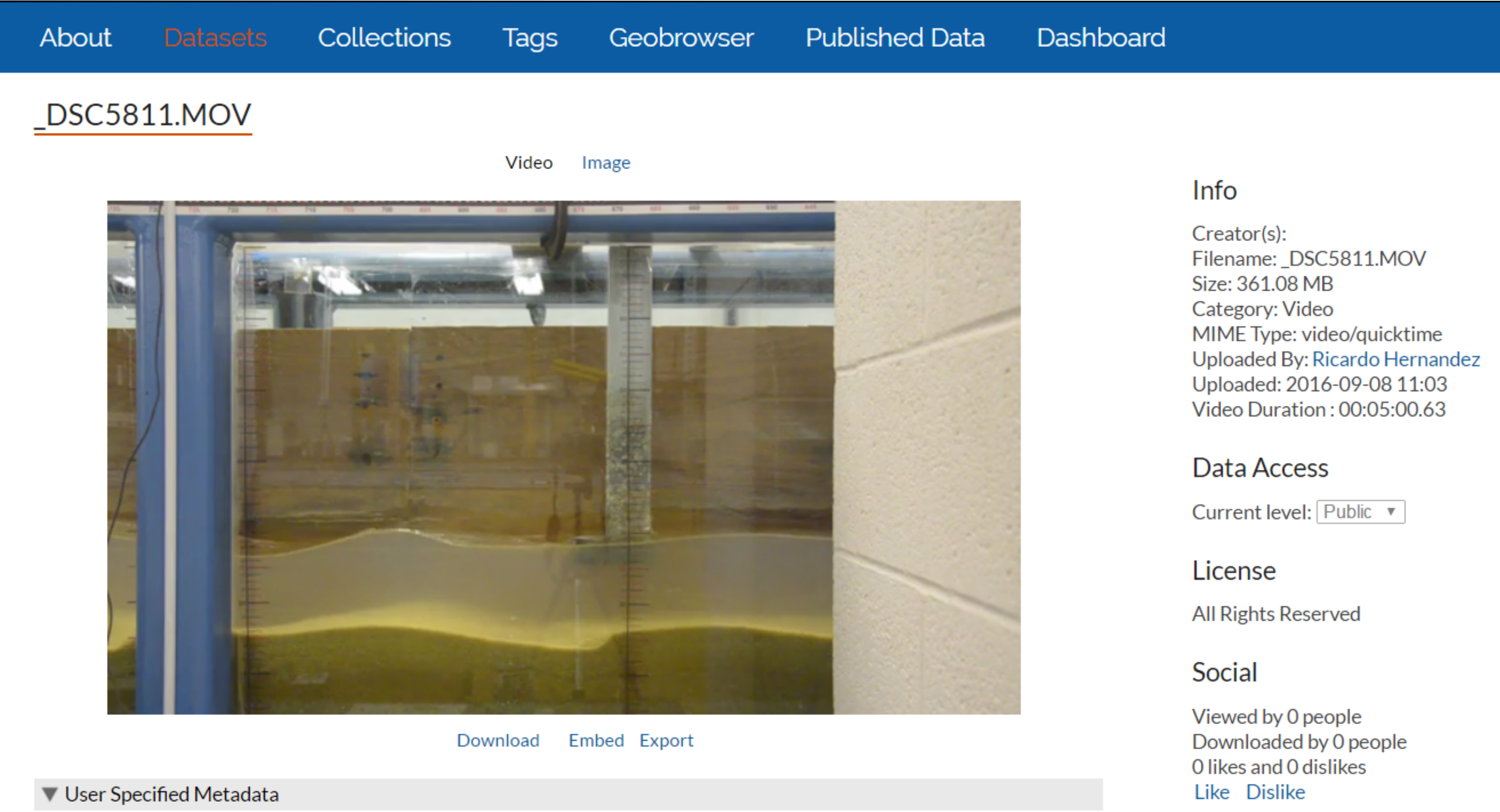
End-to-end data services for managing, sharing, curating, and publishing data

- Solutions for big data
  - Can take up to 100s of GB per project
  - SEAD desktop uploader
  - Large file preview
- Solutions for curating
  - Integration with ORCiD
  - Proof and staging areas before formal publication
  - Data publication with persistent, citable identifiers
  - Published data included in DataOne index
  - Procedure for publishing subsequent versions

SEAD Desktop Uploader

- can manage ~100,000 file uploads
- command-line java tool
- sends over whole directory structure
- keeps track of what is already uploaded,
  so updates can scan and just upload new files

SEAD: large file preview



_DSC5811.MOV

Info

Creator(s):
Filename: _DSC5811.MOV
Size: 361.08 MB
Category: Video
MIME Type: video/quicktime
Uploaded By: Ricardo Hernandez
Uploaded: 2016-09-08 11:03
Video Duration : 00:05:00.63

Data Access

Current level: Public ▼

License

All Rights Reserved

Social

Viewed by 0 people
Downloaded by 0 people
0 likes and 0 dislikes

USGS ScienceBase, sciencebase.gov

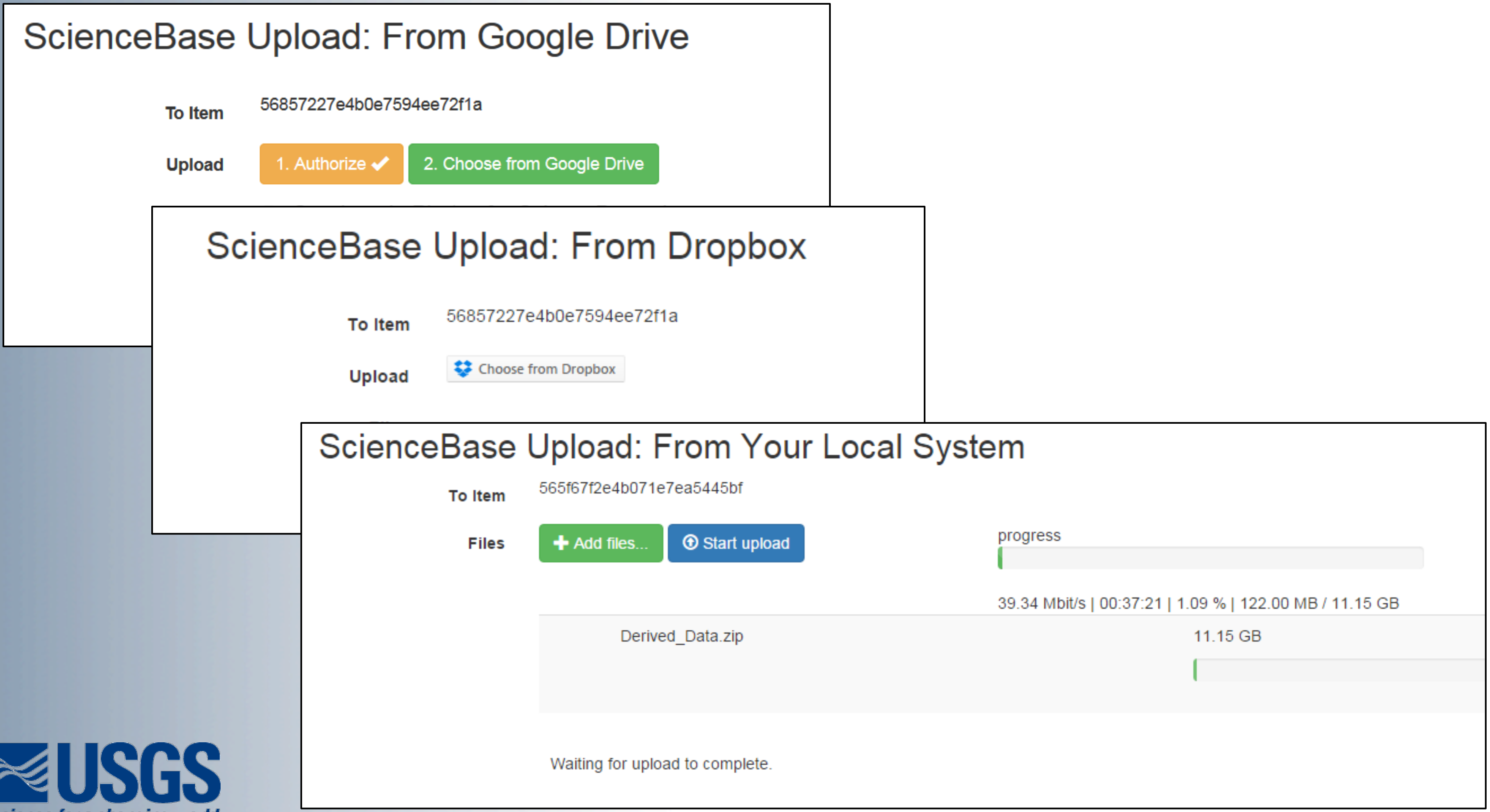
A collaborative scientific data and information management platform

- Solutions for big data:
  - Can take 100s of GB per project
  - Large file uploader and downloader
  - Embedded video previews, linked with YouTube

- Solutions for curating:
  - Persistent, citable identifier
  - Robust metadata requirements for Data Releases
  - Included in USGS Science Data Catalog and data.gov
  - APIs and extensions add value to data

ScienceBase Large File Uploader

Allows users to upload files from Google Drive, Dropbox, and local systems. Up to ~12 GB per file.



ScienceBase Upload: From Google Drive

To Item 56857227e4b0e7594ee72f1a

Upload

ScienceBase Upload: From Dropbox

To Item 56857227e4b0e7594ee72f1a

Upload

ScienceBase Upload: From Your Local System

To Item 565f67f2e4b071e7ea5445bf

Files

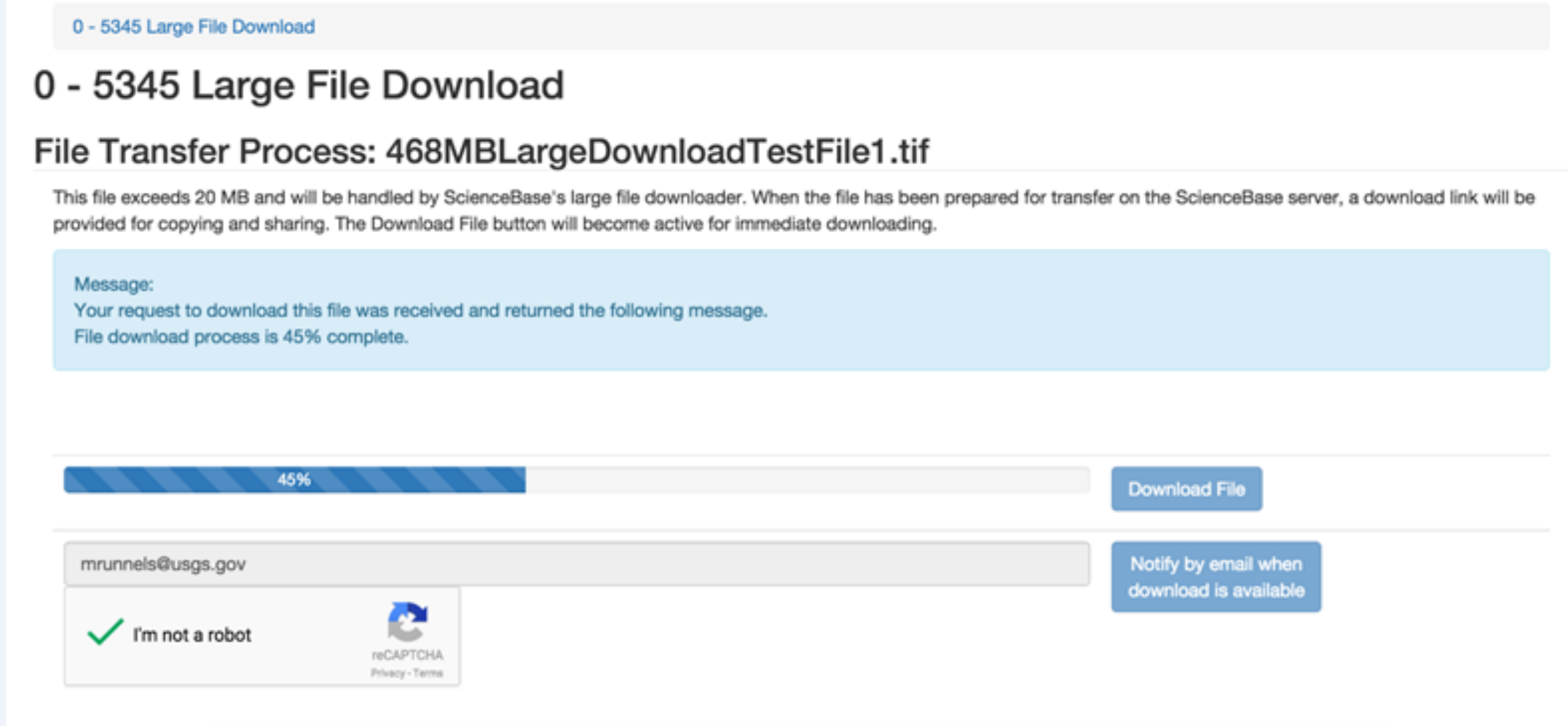
progress

39.34 Mbit/s | 00:37:21 | 1.09 % | 122.00 MB / 11.15 GB

Derived_Data.zip 11.15 GB

Waiting for upload to complete.

ScienceBase Large File Downloads



0 - 5345 Large File Download

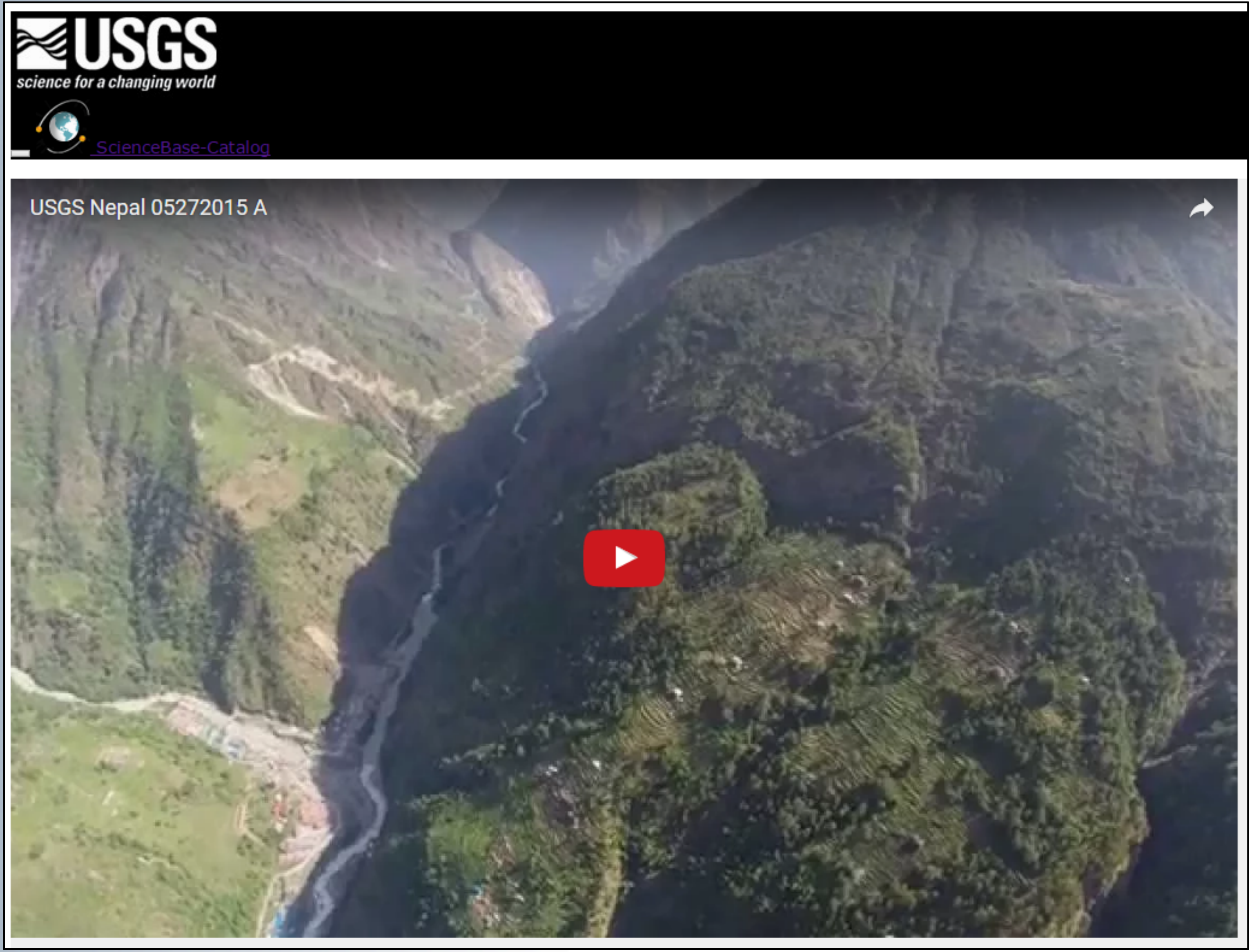
0 - 5345 Large File Download

File Transfer Process: 468MBLargeDownloadTestFile1.tif

This file exceeds 20 MB and will be handled by ScienceBase's large file downloader. When the file has been prepared for transfer on the ScienceBase server, a download link will be provided for copying and sharing. The Download File button will become active for immediate downloading.

Message:
Your request to download this file was received and returned the following message.
File download process is 45% complete.

ScienceBase Preview of large files with YouTube Integration



The potential of data in the long tail

Research Data Alliance Sept 2016 Session:
Making Small Data BIG

AGU 2016 Earth and Space Science Informatics Session:
BIG Value of Small Data

"Like pieces of a puzzle that create a picture when put together correctly, small data, when properly curated and aggregated, can reveal large-scale temporal and spatial patterns that lead to major new scientific discoveries.” *[AGU Session Abstract]*

Summary

- Datasets in the heterogeneous Long Tail are increasing in size
- We are starting to find solutions for curating and publishing this Big Data
- Communities of Practice help to find and develop solutions
- Proper curation of (Big) long-tail datasets has huge scientific potential
- We can learn from different disciplines – comparing challenges and solutions

Discuss challenges and issues like this with your colleagues in the GSA Geoinformatics Division