# Collaborative Research Data Repository with Hyrax

Jim Halliday and Nabeela Jaffer Indiana University, University of Michigan Samvera Connect 2019 October 24, 2019

## A slight shift in focus...

... and more questions than answers!

## So what is research data, anyway?

"...any information collected, stored, and processed to produce and validate original research results" (https://libguides.macalester.edu/data1)

"...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (https://www.lib.ncsu.edu/data-management/define)

"...materials generated or collected during the course of conducting research" (https://www.neh.gov/sites/default/files/2018-06/data\_management\_plans\_2018.pdf)

## So what is research data, anyway?

- There is no one definitive answer
- Research data is NOT limited to the sciences.
- It is not always just 'numbers'.
- We want to preserve and (sometimes) present it!

## Research Data File Types

- Simple text files (CSV)
- Complex binary data (the infamous .dat files)
- Images
- Sound
- Zip files
- Absolutely NO standards here!

## Research Data File Types

#### Even Entire Disk Images!



- Jetstream is a cloud computing environment that provides virtual machines to researchers for data storage and computation.
- At IU, researchers work in Jetstream, and when finished the entire VM is tarred and backed up into our DSpace repository.

#### Specific Challenges for Digital Repositories

#### File Size

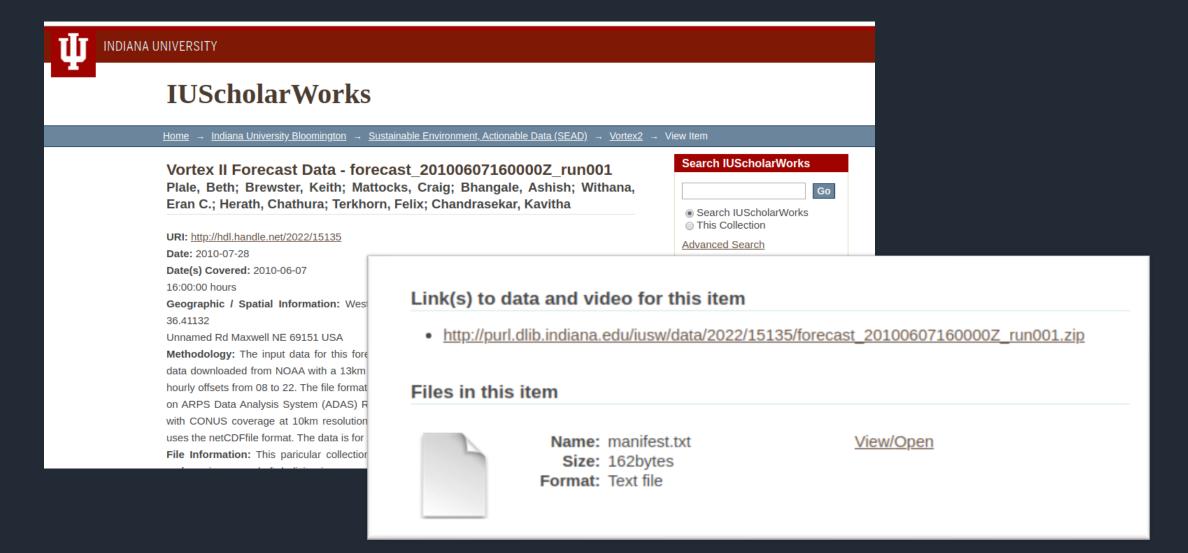
- Huge variety of file sizes, from tiny text files to multiple terabytes and beyond
- Huge files require large amounts of storage
- Researchers sometimes have trouble uploading large files
- IU has the Scholarly Data Archive (SDA) tape storage system, which helps but brings its own set of problems

#### Specific Challenges for Digital Repositories

#### Non-uniform file types and metadata

- Researchers use metadata in a large number of ways
- Consistent standards are emerging for some types of data
- Variety of file types makes this challenging

## IU Research Data in DSpace

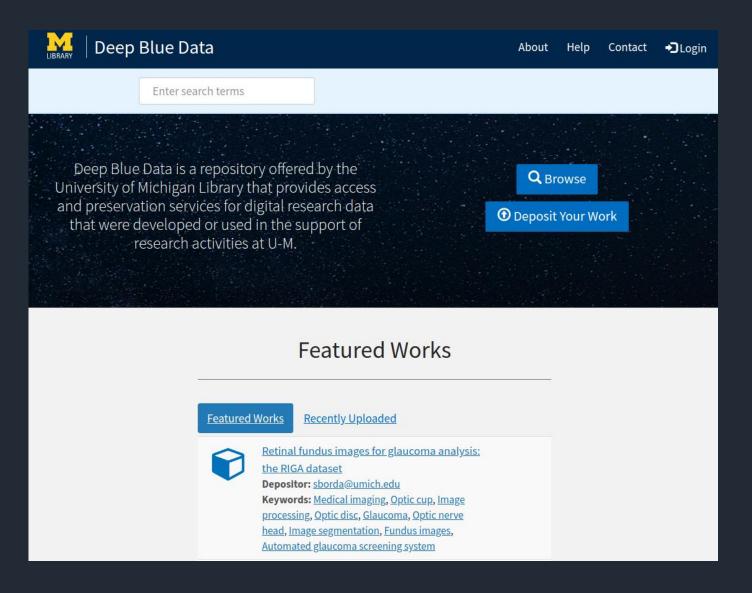


#### IU Research Data in DSpace

#### Problems with this approach

- Data content mixed in with PDF's and other non-research data
- Metadata is not data-centric
- Relying on tape causes delays in downloading data
- Large zip files have to be downloaded entirely before use

## Deep Blue Data – A Better Approach



- Hyrax based
- A data-specific repository

## Moving Towards Chimera

Chimera is a new effort to create a generic data research repository based on Hyrax.

It is based on Deep Blue Data's code base.

Available at

<u> https://github.com/samvera-labs/chimera</u>



## Chimera Challenges

Coming up with a shared, generic data repository proved to be quite challenging.

There have been issues with:

- Branding
- Authentication
- DOI Creation
- Permissions

Toggling between institution-specific features is messy!

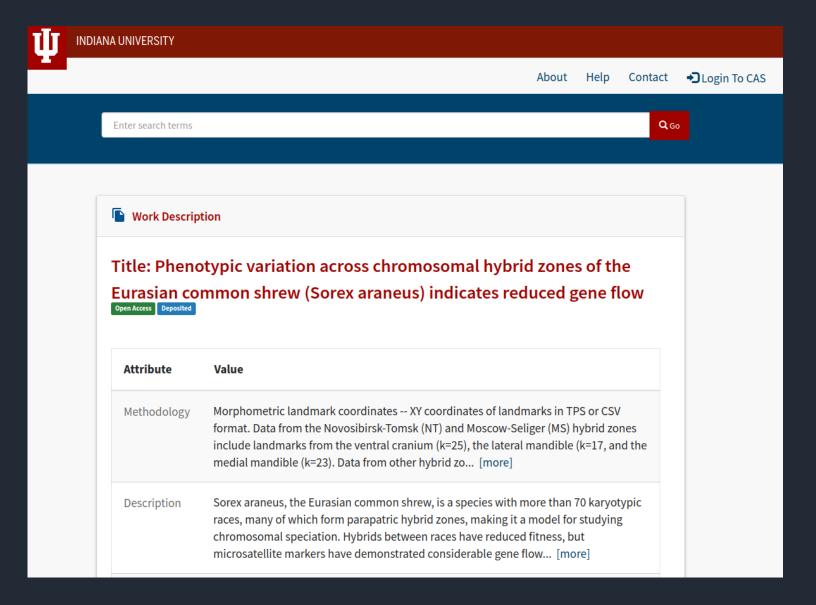
#### IU Data CORE

Date CORE (the IU Data Catalog and Open Repository) is IU's upcoming implementation of Chimera.

It will provide IU-specific features such as:

- IU CAS Login
- Appropriate branding
- Permission restrictions by campus
- DOI Creation

#### IU Data CORE



#### IU Data CORE

All the existing data sets currently in DSpace (around 200) will be migrated to Data CORE.

The metadata will be mapped appropriately from Dublin Core.

Many other new data sets are in the wings, waiting to be ingested.

#### What's Next?

- Put IU Data Core into a pilot phase and migrate DSpace data content.
- Push IU-specific changes back into Chimera.
- Improve vanilla Chimera so it can be used out of the box with configurable features.

## Thank you!