

# 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](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

 INDIANA UNIVERSITY

## IUScholarWorks

Home → Indiana University Bloomington → Sustainable Environment, Actionable Data (SEAD) → Vortex2 → View Item

### Vortex II Forecast Data - forecast\_20100607160000Z\_run001

Plale, Beth; Brewster, Keith; Mattocks, Craig; Bhangale, Ashish; Withana, Eran C.; Herath, Chathura; Terkhorn, Felix; Chandrasekar, Kavitha

**URI:** <http://hdl.handle.net/2022/15135>  
**Date:** 2010-07-28  
**Date(s) Covered:** 2010-06-07  
16:00:00 hours  
**Geographic / Spatial Information:** West 36.41132  
Unnamed Rd Maxwell NE 69151 USA  
**Methodology:** The input data for this forecast data downloaded from NOAA with a 13km hourly offsets from 08 to 22. The file format on ARPS Data Analysis System (ADAS) R with CONUS coverage at 10km resolution uses the netCDFfile format. The data is for  
**File Information:** This particular collection

**Search IUScholarWorks**

Search IUScholarWorks  
 This Collection

[Advanced Search](#)


---

### Link(s) to data and video for this item

- [http://purl.dlib.indiana.edu/iusw/data/2022/15135/forecast\\_20100607160000Z\\_run001.zip](http://purl.dlib.indiana.edu/iusw/data/2022/15135/forecast_20100607160000Z_run001.zip)

---

### Files in this item

	<b>Name:</b> manifest.txt <b>Size:</b> 162bytes <b>Format:</b> Text file	<a href="#">View/Open</a>
--	--	---------------------------

# 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

The screenshot shows the Deep Blue Data website. At the top left is the University of Michigan Library logo. The navigation bar includes 'About', 'Help', 'Contact', and 'Login'. A search bar is present with the placeholder text 'Enter search terms'. The main content area has a dark background with a starry pattern. It contains a paragraph describing the repository: 'Deep Blue Data is a repository offered by the University of Michigan Library that provides access and preservation services for digital research data that were developed or used in the support of research activities at U-M.' To the right of this text are two blue buttons: 'Browse' with a magnifying glass icon and 'Deposit Your Work' with an upload icon. Below this is a 'Featured Works' section with two tabs: 'Featured Works' (active) and 'Recently Uploaded'. The first featured work is titled 'Retinal fundus images for glaucoma analysis: the RIGA dataset' and includes the depositor's email and a list of keywords.

**M LIBRARY** | Deep Blue Data About Help Contact Login

Enter search terms


Deep Blue Data is a repository offered by the University of Michigan Library that provides access and preservation services for digital research data that were developed or used in the support of research activities at U-M.

[Browse](#)

[Deposit Your Work](#)

## Featured Works

[Featured Works](#) [Recently Uploaded](#)

 [Retinal fundus images for glaucoma analysis: the RIGA dataset](#)  
**Depositor:** [sborda@umich.edu](mailto:sborda@umich.edu)  
**Keywords:** [Medical imaging](#), [Optic cup](#), [Image processing](#), [Optic disc](#), [Glaucoma](#), [Optic nerve head](#), [Image segmentation](#), [Fundus images](#), [Automated glaucoma screening system](#)

- 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

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



# 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



INDIANA UNIVERSITY

[About](#) [Help](#) [Contact](#) [Login To CAS](#)

Enter search terms

Go

## Work Description

### Title: Phenotypic variation across chromosomal hybrid zones of the Eurasian common shrew (*Sorex araneus*) indicates reduced gene flow

[Open Access](#) [Deposited](#)

Attribute	Value
Methodology	Morphometric landmark coordinates -- XY coordinates of landmarks in TPS or CSV format. Data from the Novosibirsk-Tomsk (NT) and Moscow-Seliger (MS) hybrid zones include landmarks from the ventral cranium (k=25), the lateral mandible (k=17, and the medial mandible (k=23). Data from other hybrid zo... <a href="#">[more]</a>
Description	<i>Sorex araneus</i> , the Eurasian common shrew, is a species with more than 70 karyotypic races, many of which form parapatric hybrid zones, making it a model for studying chromosomal speciation. Hybrids between races have reduced fitness, but microsatellite markers have demonstrated considerable gene flow... <a href="#">[more]</a>

# 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!

Jim Halliday, [jhallida@indiana.edu](mailto:jhallida@indiana.edu)