DLF 2018

# Putting theory into practice: Lessons from the Data Curation Network


***Lisa Johnston***
*University of Minnesota*

***Jake Carlson***
*University of Michigan*

***Wendy Kozlowski***
*Cornell University*

***Robert Olendorf***
*Penn State University*


*October 16, 2018*

Data Curation Network

# Outline of Talk

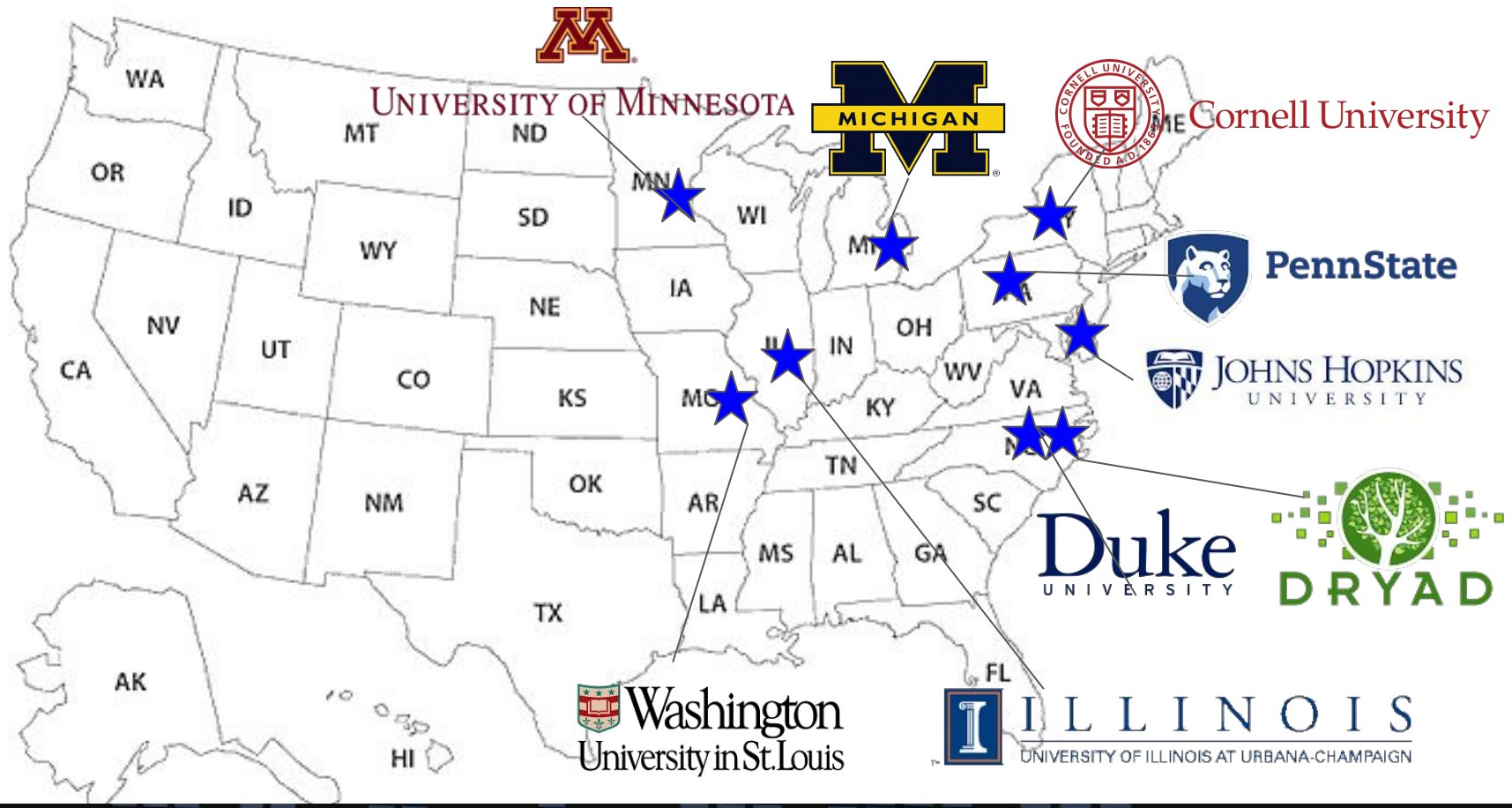
1. 5m - Introduction to the DCN - Lisa
2. 20m - What we mean by data curation - Rob and Wendy
3. 10m - Building the DCN curator community - Jake
4. 5m - Building the broader data curation community - Jake
5. 5 min - Next steps for the DCN - Lisa
6. 10m for q&a

***Researchers are faced with a growing number of requirements (and incentives) to ethically share their research data.***

*Well curated data are more valuable.*

***The skills and expertise required to curate data cannot be fully automated nor reasonably be provided by a few experts siloed at single institutions.***

The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.




Data Curation Network

# DCN Progress to Date

Alfred P. Sloan FOUNDATION

**2016 – 2017**

Planning phase with six institutions

**July 2017**

DCN Model released for comment

**2018-2021**

Three-year implementation phase with eight institutions

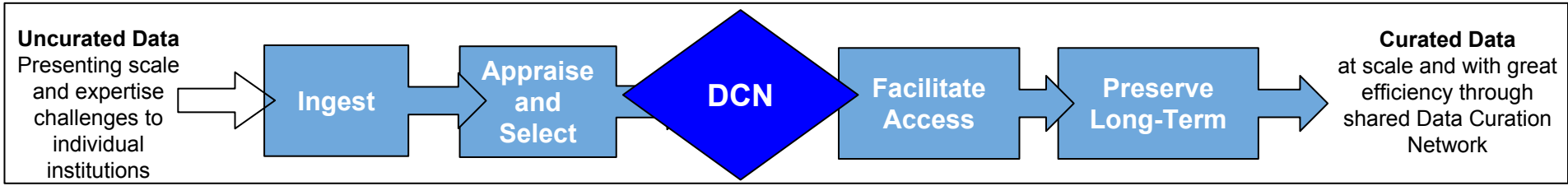
**2020-Beyond**

Open the network to new members and end-users

http://DataCurationNetwork.org

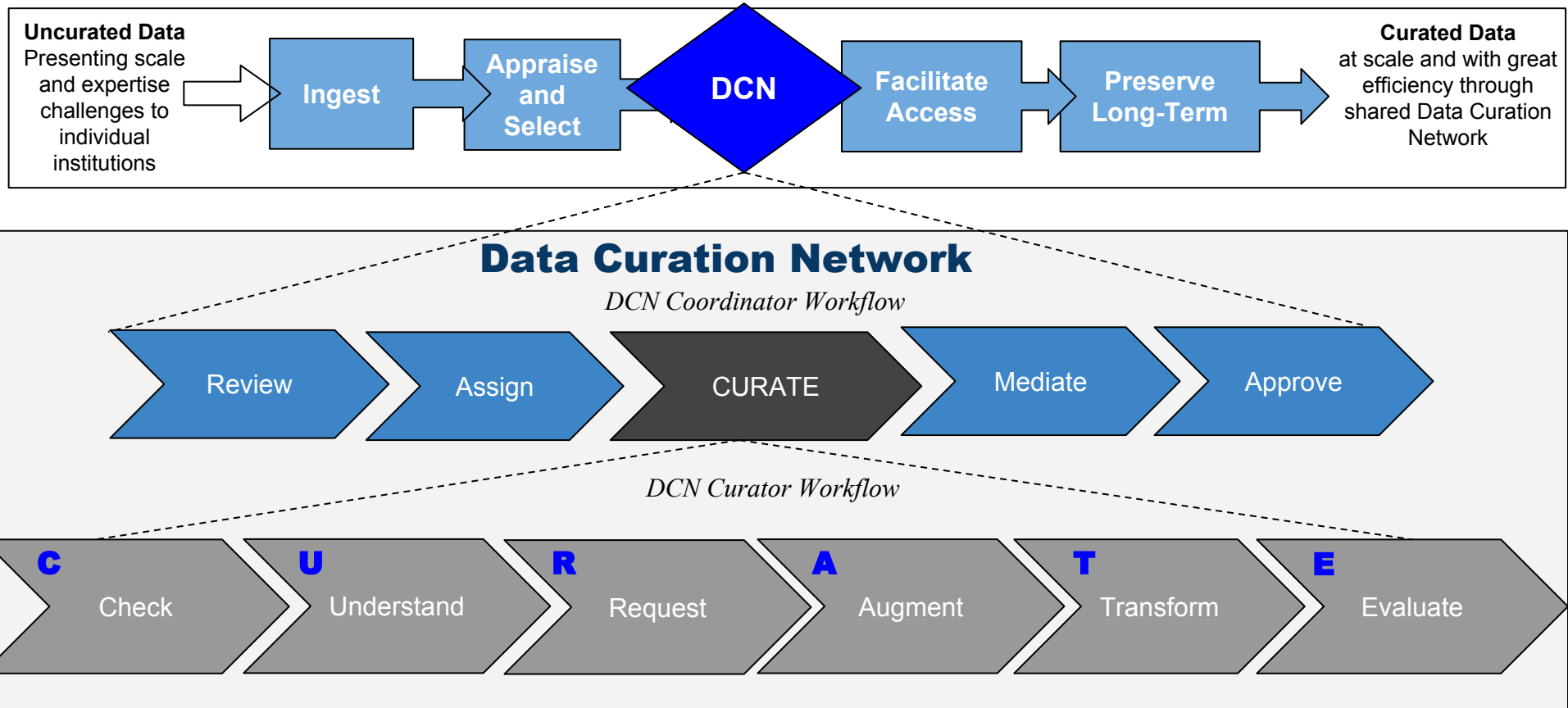
# What do we really mean by data curation?

# DCN Workflow




- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)

# DCN Workflow



# CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not...
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.

# CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not...
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.
- **D** **Document** curation activities (accessioning, provenance, workflow)

# DCN Checklists

**Table A1.** Draft checklist of DCN CURATE steps and FAIRness scorecard

| CURATE Actions | Curation Checklist |
|---|---|
| **Check** data files and read documentation<br>▪ Review the content of the data files (e.g., open and run the files or code).<br>▪ Verify all metadata provided by the author and review the available documentation. | ☐ Files open as expected<br>☐ Issues _____<br>☐ Code runs as expected<br>☐ Produces minor<br>☐ Does not run an<br>many errors<br>☐ Metadata quality is rich,<br>complete<br>☐ Metadata has iss<br>☐ Documentation Type (*ci*<br>Readme / Codebook / D<br>Other: _____<br>☐ Missing/None<br>☐ Needs work |
| **Understand** the data (or try to)<br>▪ Check for quality assurance and usability issues such as missing | *Varies based on file formats and s*<br>*example….* |

**Evaluate** and rate the overall data record for FAIRness.[2]
- Score the dataset and recommend ways to increase the FAIRness of the data and become "DCN approved."

Findable -
- ☐ Metadata exceeds author/ title/ date,
- ☐ Unique PID (DOI, Handle, PURL, etc.).
- ☐ Discoverable via web search engines like Google.

Accessible -
- ☐ Retrievable via a standard protocol (e.g., HTTP).
- ☐ Free, open (e.g., download link).

Interoperable -
- ☐ Metadata formatted in a standard schema (e.g., Dublin Core).
- ☐ Metadata provided in machine-readable format (OAI feed).

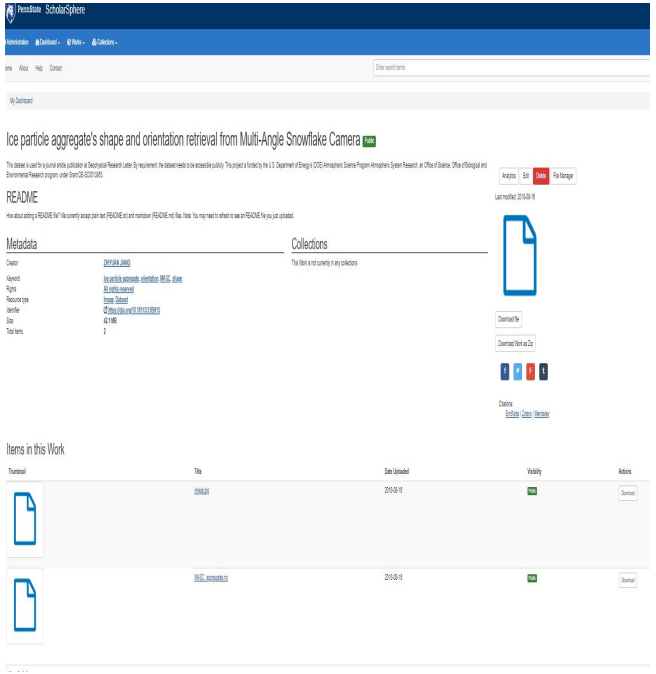
Reusable -
- ☐ Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author.
- ☐ Clear indicators of who created, owns, and stewards the data.
- ☐ Data are released with clear data usage terms (e.g., a CC License).

[1] Format Recommendations, http://guides.library.cornell.edu/ecommons/formats
[2] Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

http://z.umn.edu/curate

# Most Data Is Initially Less Than Good

- Often only data
- Code doesn't run (if present)
- No description
- No README
- No Data Dictionary
- Poorly Name




https://scholarsphere.psu.edu/concern/generic_works/gxd07gs54v

# Ice particle aggregate's shape and orientation retrieval from Multi-Angle Snowflake Camera Public

This dataset is used for a journal article publication at Geophysical Research Letter. By requirement, the dataset needs to be accessible publicly. This project is funded by the U.S. Department of Energy's (DOE) Atmospheric Science Program Atmospheric System Research, an Office of Science, Office of Biological and Environmental Research program, under Grant DE-SC0013953.

## README

How about adding a README file? We currently accept plain text (README.txt) and markdown (README.md) files. Note: You may need to refresh to see an README file you just uploaded.

## Metadata

| | |
|---|---|
| Creator | ZHIYUAN JIANG |
| Keyword | Ice particle aggregate, orientation, MASC, shape |
| Rights | All rights reserved |
| Resource type | Image, Dataset |
| Identifier | https://doi.org/10.18113/S1BH15 |
| Size | 42.1 MB |
| Total items | 2 |

## Collections

This Work is not currently in any collections.

Analytics

Last modified: 2018-08-16

Download file

Download Work as Zip

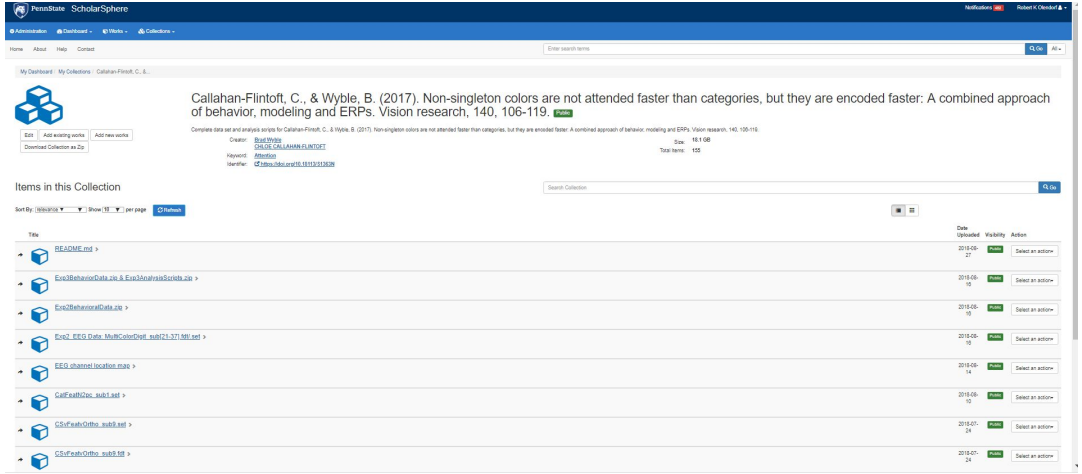
Citations:
EndNote | Zotero | Mendeley

## Items in this Work

| Thumbnail | Title | Date Uploaded | Visibility | Actions |
|---|---|---|---|---|
|  | image.zip | 2018-08-16 | Public | Download |
|  | MASC_aggregates.nc | 2018-08-16 | Public | Download |

# It Takes A Lot Of Effort To Improve The Data A Little

- Specialized File Types Are Common
- Domain Experience Helps
- It Takes A Lot of “Collaboration” With the Researcher(s)
- It Takes Time (Weeks)
- Researchers Can Get Frustrated
- Researchers Push Back
- Incremental Improvements Are Expected




https://scholarsphere.psu.edu/collections/79407x18h

# Callahan-Flintoft, C., & Wyble, B. (2017). Non-singleton colors are not attended faster than categories, but they are encoded faster: A combined approach of behavior, modeling and ERPs. Vision research, 140, 106-119. Public

Edit | Add existing works | Add new works | Download Collection as Zip

Complete data set and analysis scripts for Callahan-Flintoft, C., & Wyble, B. (2017). Non-singleton colors are not attended faster than categories, but they are encoded faster: A combined approach of behavior, modeling and ERPs. Vision research, 140, 106-119.

Creator: Brad Wyble
CHLOE CALLAHAN-FLINTOFT

Keyword: Attention

Identifier: https://doi.org/10.18113/S1363N

Size: 18.1 GB

Total Items: 155

## Items in this Collection

Search Collection | Go

Sort By: relevance ▼ Show 10 per page | Refresh

| Title | Date Uploaded | Visibility | Action |
|---|---|---|---|
| README.md | 2018-08-27 | Public | Select an action |
| Exp3BehaviorData.zip & Exp3AnalysisScripts.zip | 2018-08-16 | Public | Select an action |
| Exp2BehavioralData.zip | 2018-08-16 | Public | Select an action |
| Exp2_EEG Data: MultiColorDigit_sub[21-37].fdt/.set | 2018-08-16 | Public | Select an action |
| EEG channel location map | 2018-08-14 | Public | Select an action |
| CatFeatN2pc_sub1.set | 2018- | Public | Select an action |

# The Benefits Can Be Great For Curated Data

- More Reuse of Data
- More Trust of Data
- Improved Impact of Research
- Improved Job Prospects for Researchers
- Improved Reputation of Library

https://scholarsphere.psu.edu/collections/jw827b80n

# README

## EEG and Behavioral Data and Analysis Scripts

### Creators


- Chloe Callahan-Flintoft czc213@psu.edu, Penn State University
- Brad Wyble, Penn State University


### Description

This collection includes EEG data, behavioral data, and analysis scripts for published research:

> Callahan-Flintoft, C., & Wyble, B. (2017). "Non-singleton colors are not attended faster than categories, but they are encoded faster: A combined approach of behavior, modeling and ERPs". *Vision Research*, 140, 106-119. https://doi.org/10.1016/j.visres.2017.06.013

The data for this paper was collected between 2014 and 2017.

### Usage

Analysis scripts have the following dependencies:
- MATLAB 2012 with the following extensions:
- Psychophysics Toolbox-3
- EEGlab (v14.1.1b)

### File Manifest

- `CSvFeatvOrtho_sub[1-16].set` & `CSvFeatvOrtho_sub[1-16].fdt`
  - EEG data files for Experiment 3: `.set` and `.fdt` formats returned by EEGlab when you enter the raw datafile in `.cnt` format
- `MultiColorDigit_sub[1-37].set` & `MultiColorDigit_sub[1-37].fdt`
  - EEG data files for Experiment 2
- `CatFeatN2pc_sub[1-40].set` & `CatFeatN2pc_sub[1-40].fdt`
  - EEG data files for Experiment 1
- `Exp1BehaviorData.zip`
  - `ExpSub_compact[1-40].mat` : Behavioral data files from Experiment 1
- `Exp2BehaviorData.zip`
  - `ExpSub_compact[1-37].mat` : Behavioral data files from experiment 2
- `Exp3BehaviorData.zip`
  - `ExpSub[1-16].mat` : Behavioral data files for experiment 3
- `Exp1and2AnalysisScripts.zip`

Edit | Add existing works | Add new works | Delete | Download Collection as Zip

# Soil Properties and Class 100m Grids United States Public

Three national U.S. soil point datasets — NCSS Characterization Database, the National Soil Information System (NASIS), and the Rapid Carbon Assessment (RaCA) datasets — were combined with remote sensing images and detailed conventional soil polygon maps, and used to generate complete-coverage gridded predictions of soil properties (percent organic carbon, total nitrogen, bulk density, pH, and percent sand and clay) and classes (taxonomic great group and particle size in the control section) for the Conterminous U.S. Soil covariate layers included: DEM-based derivatives, long-term MODIS EVI seasonal images, MODIS cloud fractions, and temperature images per month, PRISM climatic datasets of precipitation, temperature, and vapor pressure deficit, and bioclimatic indicators, Landsat (cloud free) NIR, SWIR bands, gamma radiometric images, geological surface classes, land cover classes, globally produced predictions of soil properties (SoilGrids250m), and the SSURGO parent material and drainage maps. The soil property and class models were built within a high-performance computing system using parallelized random forest and gradient boosting algorithms. Predictions were generated at 100 meter spatial resolution for 7 standard soil depths (0, 5, 15, 30, 60, 100 and 200 cm) for soil properties and as probabilities per soil class. Please refer to the README file for information on use and citation.


Creator: Tomislav Hengl
Sharon Waltman
Skye Wills
James Thompson
Colby Brungard
Amanda Ramcharan
Travis Nauman


Keyword: Soil Properties
Soil
Landsat
Soil Class
Soil Characteristics


Identifier: https://doi.org/10.18113/S1KW2H


Related URL: https://github.com/aramcharan/US_SoilGrids100m

Size: 52.7 GB

Total Items: 454

## Items in this Collection

Search Collection | Go

| Title | Uploaded | Visibility | Action |
|---|---|---|---|
| SoilGrids_USA48_gSSURGO_drainage.csv | 2017-07-24 | Public | Select an action |
| PMTGSS7_f.tif | 2017-07-24 | Public | Select an action |
| SoilGrids_USA48_gSSURGO_pmaterial.csv | 2017-07-24 | Public | Select an action |
| DRNGSS7_f.tif | 2017-07-24 | Public | Select an action |
| TAXgg_Xerorthents_100m.tif | 2017-05-11 | Public | Select an action |
| TAXgg_Xerumbrepts_100m.tif | 2017-05-11 | Public | Select an action |
| TAXgg_Xerofluvents_100m.tif | 2017-05-11 | Public | Select an action |
| TAXgg_Xeropsamments_100m.tif | 2017-05-11 | Public | Select an action |
| TAXgg_Xerochrepts_100m.tif | 2017-05-11 | Public | Select an action |
| TAXgg_Vitricryands_100m.tif | 2017-05-11 | Public | Select an action |

# README Public

## README

### Soil Property and Class Maps of the Conterminous US at 100 meter Spatial Resolution

Summary: These are results of spatial predictions of soil property and soil classes for the conterminous U.S. at 100 m spatial resolution. Three national U.S. soil point datasets ��? NCSS Characterization Database, the National Soil Information System (NASIS), and the Rapid Carbon Assessment (RaCA) datasets ��? were combined with remote sensing images and detailed conventional soil polygon maps, and used to generate complete-coverage gridded predictions of soil properties (percent organic carbon, total nitrogen, bulk density, pH, and percent sand and clay) and classes (taxonomic great group and particle size in the control section). An ensemble model from the machine learning algorithms random forest and gradient boosting, as implemented in R packages ranger and xgboost, were used to generate spatial predictions. The model validation results indicate an average classification accuracy of 60% for great groups, and 66% for texture classes; for soil properties R-square at validation points ranged from 62% for total nitrogen to 87% for pH. This hybrid "SoilGrids+" modelling system that incorporates remote sensing data, global and local predictions of soil properties, traditional soil polygon maps, and machine learning, opens a possibility for combining traditional (soil survey data) with state-of-the-art Machine Learning technology, with an objective to make soil data more accurate, easier to update, more accessible, and easier to use.


#### Manifest

The maps for download are all in the collection individually. The zip file contains all the anlayses code used for creating the maps.

#### Usage

To regenerate the maps run `Soil_Property_and_Class_Models_Codes/m_NASIS_100m.R`. The remaining R files are called by that script. The maps will be output in the results directory.

#### Citation

To cite the relevant article cite:

- Ramcharan A.,Hengl T., Nauman T., Brungard C., Waltman S., Wills S., Thompson J. (2017) **Soil Property and Class Maps of the Conterminous US at 100 meter Spatial Resolution based on a Compilation of National Soil Point Observations and Machine Learning.** Submitted to Soil Science Society of America Journal.

To cite the data directly use:
* Ramcharan A.,Hengl T., Nauman T., Brungard C., Waltman S., Wills S., Thompson J. (2017) **Soil Properties and Class 100m Grids United States**. https://doi.org/10.18113/S1KW2H

#### Disclaimer

These are results of spatial predictions based on using Machine Learning algorithms attached to the above-listed paper and hence some errors and artifacts are still possible. We aim at updating these maps regularly i.e. as the new training / point data arrives.

## Metadata

| | |
|---|---|
| Creator | Amanda Ramcharan, Tomislav Hengle, Travis Nauman, Colby Brungard , Sharon Waltman, Skye Wills, James Thompson |
| Keyword | Land cover, Landsat, Soil Properties, Soil, Soil Characteristics, Soil Class |
| Rights | Attribution 3.0 United States |
| Resource type | Software or Program Code |
| Published Date | 2017 |
| Location | United States |
| Size | 2.99 KB |

## Collections

Soil Properties and Class 100m Grids United States

The DCN will help support cross-institutional collaborations.

# DCN Community Collaboration Example

LSA MUSEUM OF ZOOLOGY
UNIVERSITY OF MICHIGAN

## Research Feature

### The Digitizing begins...

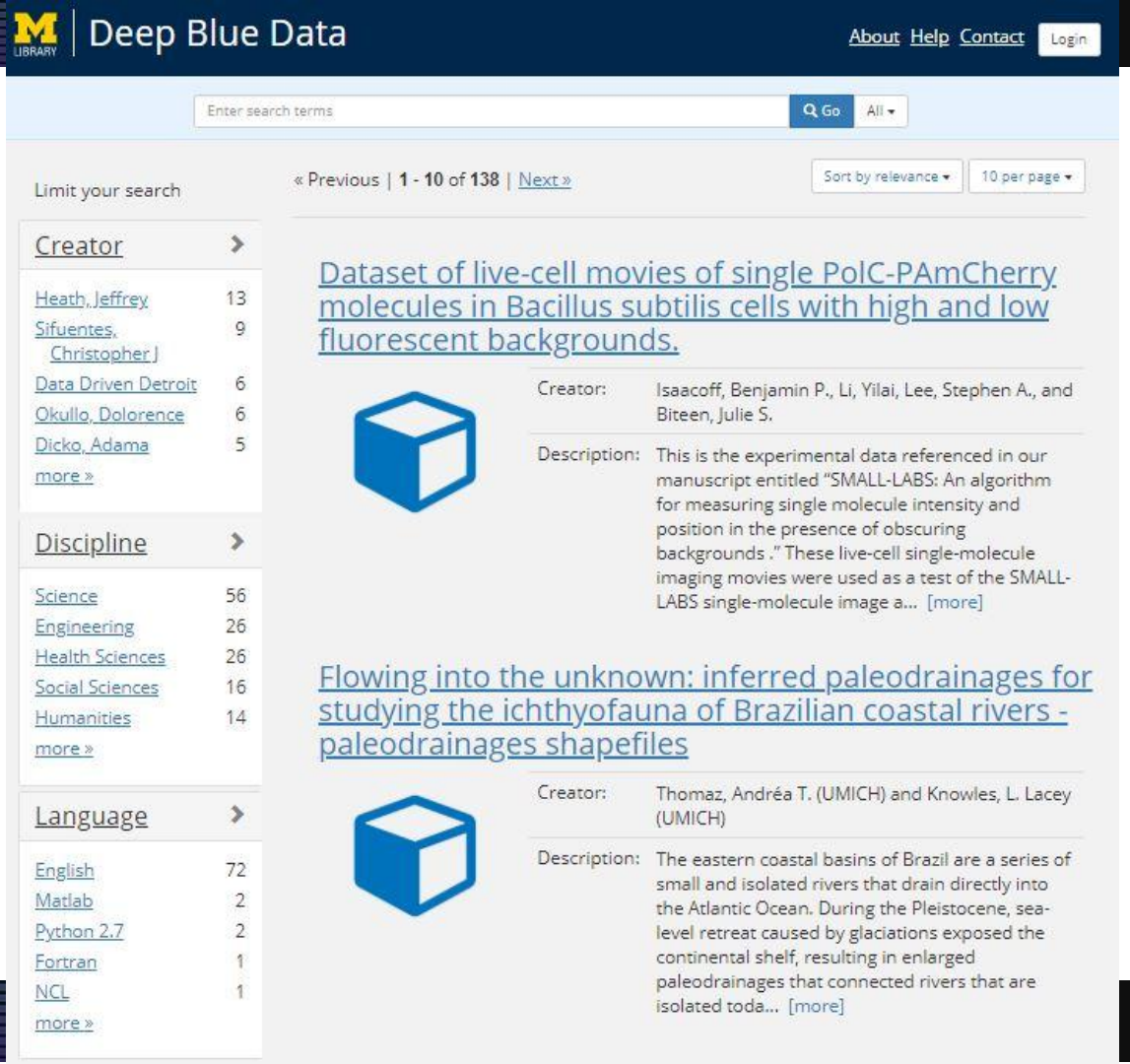
The Museum of Zoology got a massive equipment upgrade!!! A new microCT system has delivered to the Research Museums Center (RMC) and will reside adjacent to the #UMMZ collections. The microCT will allow museum researchers the ability to digitize the nearly 15 million zoological specimens, as well as other University of Michigan collections housed at the RMC. Currently, **Alison Davis Rabosky**, **Dan Rabosky**, **Cody Thompson**, and **Priscilla Tucker** are working on a National Science Foundation (NSF) grant, which will digitize ALL vertebrate genera. More projects are to come...

# CT Scan Data Workflow

1. Specimens are placed in a CT scanner which produces hundreds of images in a proprietary format
2. Images are “reconstructed” into TIFF Stacks
3. Software is used to enable 3D images of the specimen

# Deep Blue Data

University of Michigan's Institutional Data Repository



# Specimen: ummz-herps-172509, *Thamnophis rufipunctatus*

PREVIOUS BACK NEXT

VIEW SPECIMEN ON IDIGBIO

## Specimen Information

**MorphoSource Identifier:** S12127
Vouchered
**Type:** Yes
**Occurrence ID:** urn:catalog:ummz:herps:172510

**Link to specimen in home repository:** http://portal.vertnet.org/o/ummz/herps?id=172509

**Notes:** imported from iDigBio. uuid:699e11a0-83f2-44be-a85d-30117a54ab8a Occurrence ID:urn:catalog:ummz:herps:172510

**Institution:** University of Michigan Museum of Zoology, Ann Arbor, Michigan, USA

## Specimen Media

M24561, 2 files

M24581, 2 files

https://www.morphosource.org/Detail/SpecimenDetail/Show/specimen_id/12127

# Connections

DUKE UNIVERSITY LIBRARIES

iDigBio
Integrated Digitized Biocollections

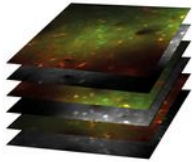
MORPHO SOURCE
BY DUKE UNIVERSITY

Metadata

M LIBRARY | Deep Blue Data

Discovery

Access and Preservation



But data curation is more than actions. The Data Curation Network will also be a community for data curators.

# Expanding our capacity to curate data

Biochemistry

Nuclear Engineering

Public Health

Archeology

Climate Science

Geospatial

TIFF

NetCDF

Text

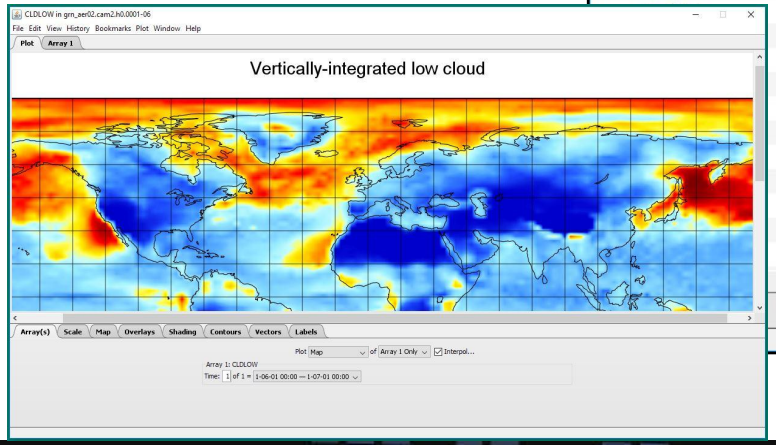
Code

CSV

PDF

Image Source:
http://www.cbsdreamteam.com/the-henry-fords-innovation-nation/

# Data Formats - NetCDF




Panoply: Sources

File Edit View History Bookmarks Plot Window Help

Create Plot | Combine Plot | Open Dataset | Remove | Remove All | Hide Info

Datasets | Catalogs | Bookmarks

| Name | Long Name | Type |
|---|---|---|
| grn_aer02.clm2.h0.0... | grn_aer02.clm2.h0.0001-06.nc | Local File |
| H2OSNO | snow depth (liquid water) | Geo2D |
| landfrac | land fraction | Geo2D |
| lat | coordinate latitude | 1D |
| lon | coordinate longitude | 1D |
| QMELT | snow melt | Geo2D |
| | time | — |
| | history time interval endpoints | 1D |

Show: All variables

**File "grn_aer02.clm2.h0.0001-06.nc"**

File type: NetCDF-3/CDM

```
netcdf file:/C:/Users/sborda/Downloads/AOD_009/grn_aer02
  dimensions:
    time = UNLIMITED;   // (1 currently)
    lat = 96;
    lon = 144;
    hist_interval = 2;
  variables:
    float H2OSNO(time=1, lat=96, lon=144);
      :long_name = "snow depth (liquid water)";
      :units = "mm";
      :cell_methods = "time: mean";
      :_FillValue = 1.0E36f; // float
      :missing_value = 1.0E36f; // float

    float QMELT(time=1, lat=96, lon=144);
```

# Developing Tools - NetCDF Primer

**Data Curation Format Profile: netCDF**
Research Data Services
University of Michigan Library

Creator(s): Sam Sciolla (ssciolla@umich.edu), Susan Borda* (sborda@umich.edu)
*denotes corresponding creator

| Core Details about netCDF | |
|---|---|
| File Extension | .nc |
| MIME Type | application/netcdf, application/x-netcdf |
| Structure | Binary, with metadata embedded in a header that can be rendered human-readable by specialized software tools |
| Versions | netCDF-4/HDF5 (post-4.0.0)<br>64-bit offset (pre-4.0.0)<br>Classic (pre-3.6.0) |
| Primary fields or areas of use | Climatology, Meteorology, Oceanography. Earth and Environment Geosciences more broadly. Also used in GIS applications. |
| Source and affiliation | NetCDF, the software and file format, is developed and managed l |

**1) Panoply**
*Website of the tool:* https://www.giss.nasa.gov/tools/panoply/.
*Version used in this profile*: 4.9.0

*What does this tool do?*
Panoply can be used to view the data stream, plot any geo-referenced and other arrays and browse the metadata.

*Who supports this tool?*
The Goddard Institute for Space Studies, a unit of the National Aeronautics and Space Administration (NASA)

| | | |
|---|---|---|
| 1. How are netCDF files organized in the dataset? | Goal | To assess the structure of the dataset as a whole and how well metadata, documentation, and file-naming conventions explain that structure |
| | Questions to answer | • How many total netCDF files are included in the dataset, and are there any zip files or directories used to divide them?<br>• Are there any clues as to the reasons or rationale for the file or directory divisions? |

http://bit.ly/NetCDFPrimer

# DCN Training Workshop

**Attendees will:**

1. Increase understanding of data curation practices and tools in various disciplines, data types, and formats
2. Share expertise and enhance curation capacity for librarians nationwide
3. Meet like-minded colleagues who are interested in building and extending curation practices at their institutions.
4. Develop a Primer to address specific challenges in curating data

DLF

October 17-18th | Spring 2019 | Fall 2019

INSTITUTE of Museum and Library SERVICES

# DCN Training Workshop

Washington University in St. Louis

RESEARCH | DATA | GIS  ABOUT US  SERVICES  RESOURCES  PROJECTS  COMMUNITY  NEWS  EVENTS

RESEARCH | DATA | GIS > DATA CURATION WORKSHOP

DATA CUR

## DATA CURATION WORKSHOP

## SLIDES AND HANDOUTS

**DATE:** DECEMBER 11 & 12, 2017

**TWEET:** #DCW2017

**LOCATION:** WASHINGTON UNIVERSITY IN ST. LOUIS, **MCMILLAN HALL,** ST. LOUIS, M

**DESCRIPTION:**

This free, 1.5 day workshop is open to all library staff and data professionals who are interested in data curation.

Participants will learn practical, hands-on treatments for data curation based on the **Data Curation Network CURATE** mod

- **C** – Check data files and read documentation;
- **U** – Understand the data (try to), if not...
- **R** – Request missing information or changes;
- **A** – Augment the submission with metadata for findability;
- **T** – Transform file formats for reuse and long-term preservation;
- **E** – Evaluate and rate the overall submission for FAIRness.

**ATTENDEES WILL COME AWAY WITH:**

1. A customized, implementable plan to enhance data curation activities at your local institution or organization,
2. Stakeholder focused talking points related to the value of data curation activities,
3. An in-depth understanding of specialized data curation practices in various disciplines, data types, and formats.

ork

# Next steps for the DCN and our plans for sustainability.

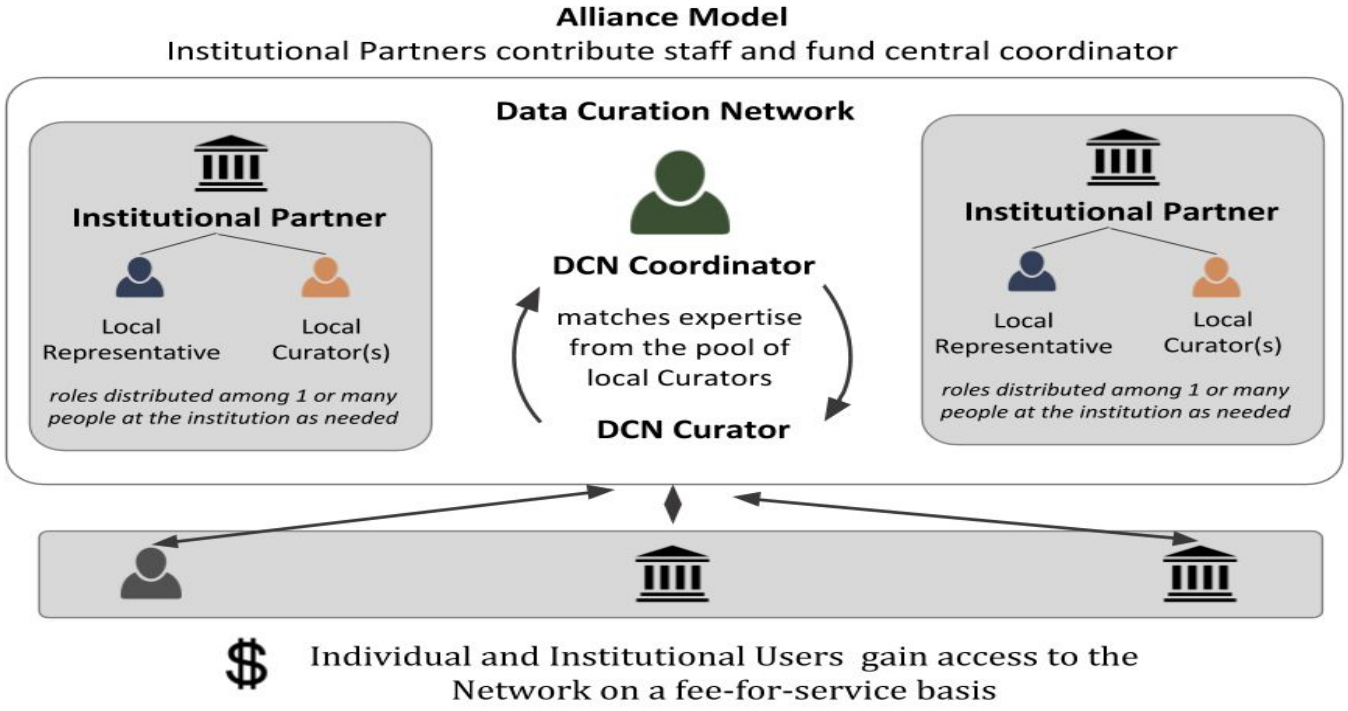
# 6-year Roadmap toward Sustainability

*Transition from planning phase to sustaining phase*

| | Year 0 | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 | Year 6 |
|---|---|---|---|---|---|---|---|
| **Support** | Sloan Grant | Grant Funded (Y1-Y2) transition to partnership model (Y3) | | | Curation-as-service (Y4-6) | | |
| **Timing** | 2016-17 | 2017-19 | | 2020-22 | | 2022-2023 | |
| **Phase** | Planning | Implementation | | Transition | | Sustaining | |
| **Partners** | 6 academic institutions | 8 academic institutions and 2 disciplinary partners | | | Recruit new partners as use and demand dictate | | |

**Mission: With a proven and appealing value-proposition, the Data Curation Network will expand into a sustainable entity that grows beyond our initial partner institutions.**

# Data Curation-as-service



# Thanks!

https://DataCurationNetwork.org

Twitter #DataCurationNetwork