SUSTAINABLE DATA AND SOFTWARE CURATION?!

From an academic library data workflows specialist/generalist data curator

LIBRARY

Susan Borda
sborda@umich.edu
@mutanthumb

NO!

And yes…

WHAT IS CURATION?

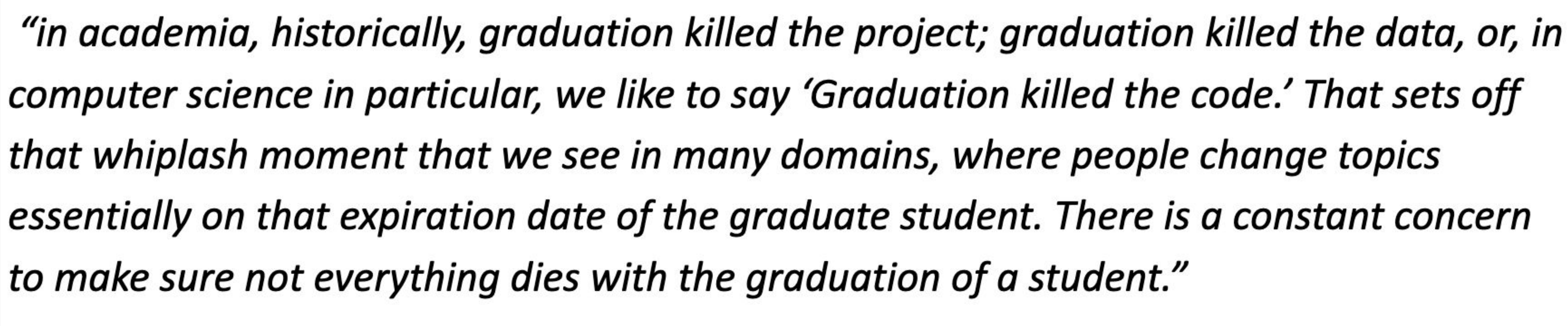
- Data curation is the process of making a dataset fit-for-use and archiveable. - A Thomer (2022)
- The active and ongoing management of data through its life-cycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation. - University of Illinois Urbana-Champaign School of Information Science – Data Curation specialization (2016) *program no longer exists

WHO WILL/CAN DO IT?

- Graduate students
- Library staff
- Other
- Researchers themselves

"Researchers are currently spending a significant portion of their own time dealing with data curation; in some cases, over 50% of their funded time. " - Mullendore GL, Mayernik MS and Schuster DC (2021)

Graduate Students?



> *"in academia, historically, graduation killed the project; graduation killed the data, or, in computer science in particular, we like to say 'Graduation killed the code.' That sets off that whiplash moment that we see in many domains, where people change topics essentially on that expiration date of the graduate student. There is a constant concern to make sure not everything dies with the graduation of a student."*

Labou, S., Otsuji, R., & Minor, D. (2021).

PEOPLE IN THE LIBRARY?

Of the 15 schools in the Data Curation Network only 3 (VT, Cornell, and UCSB) have people listed as "experts" in Simulation data and practically speaking fewer than that can actually handle data that is 1TB or more.

A cursory review of current open "Data Curator" positions (at universities) majority were for genomics or NIH based funded research others were more general

Software mentioned as a "desired skill" not necessarily required

U Mich has an opening for a "Data Curation and Research Reproducibility Specialist" person this is split position between MIDAS (the Data Science Dept on campus) and the Library.

LIBRARY/ISCHOOL PROGRAMS

- UW Milwaukee Data Curation cert: https://catalog.uwm.edu/information-studies/data-curation-graduate-certificate/#requirementstext
- IU Digital curation specialization : https://ils.indiana.edu/programs/specializations/digital-curation.html
- UT Austin data curation specialization(?): https://www.ischool.utexas.edu/programs/endorsement
- Other schools including U Mich have single classes focusing on Digital or Data Curation.

SKILLS & RESOURCES NEEDED FOR DATA/SOFTWARE CURATORS

- Troubleshooting
- Command line scripting
- MATLAB
- Time
- Access to large amounts of storage (limiting factor)
- Python & Julia (and of course Fortran)
- Server access for review and pre-ingest processing (limiting factor)

DATA AND SOFTWARE CURATOR NEEDS FROM THIS COMMUNITY

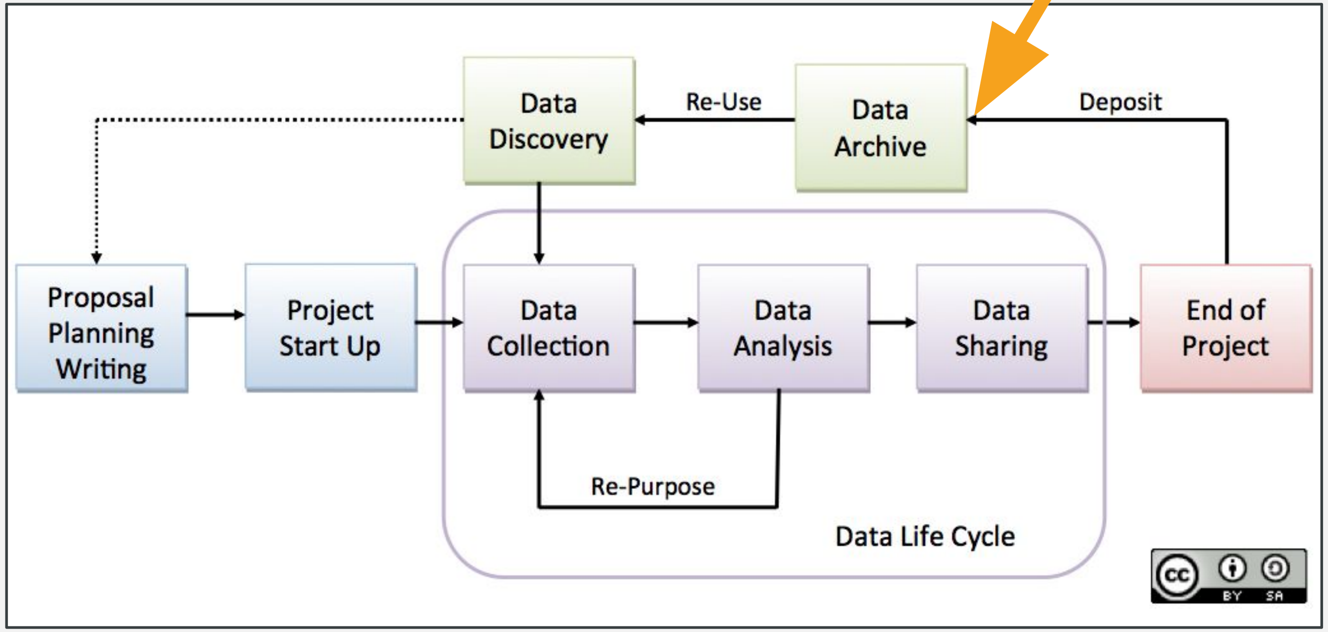
- Specific training and guidance for "generalist" curators
  - Related to netCDF, HDF5, MATLAB, Python, etc. and how they are used in Atmospheric research.
  - HPC training/understanding might be helpful as well.

- Investment in the data repository tools and infrastructure
  - Dspace, Hyrax, Dataverse, Invenio, Rucio, etc
  - (This is more for Data) Globus (or other means of transferring large files) integration
    - I have shared my Globus workflow for large datasets with Duke, Penn State, U of Arizona, U of Illinois, and the U of Alabama – Birmingham, UW – Madison, and others (via webinar).

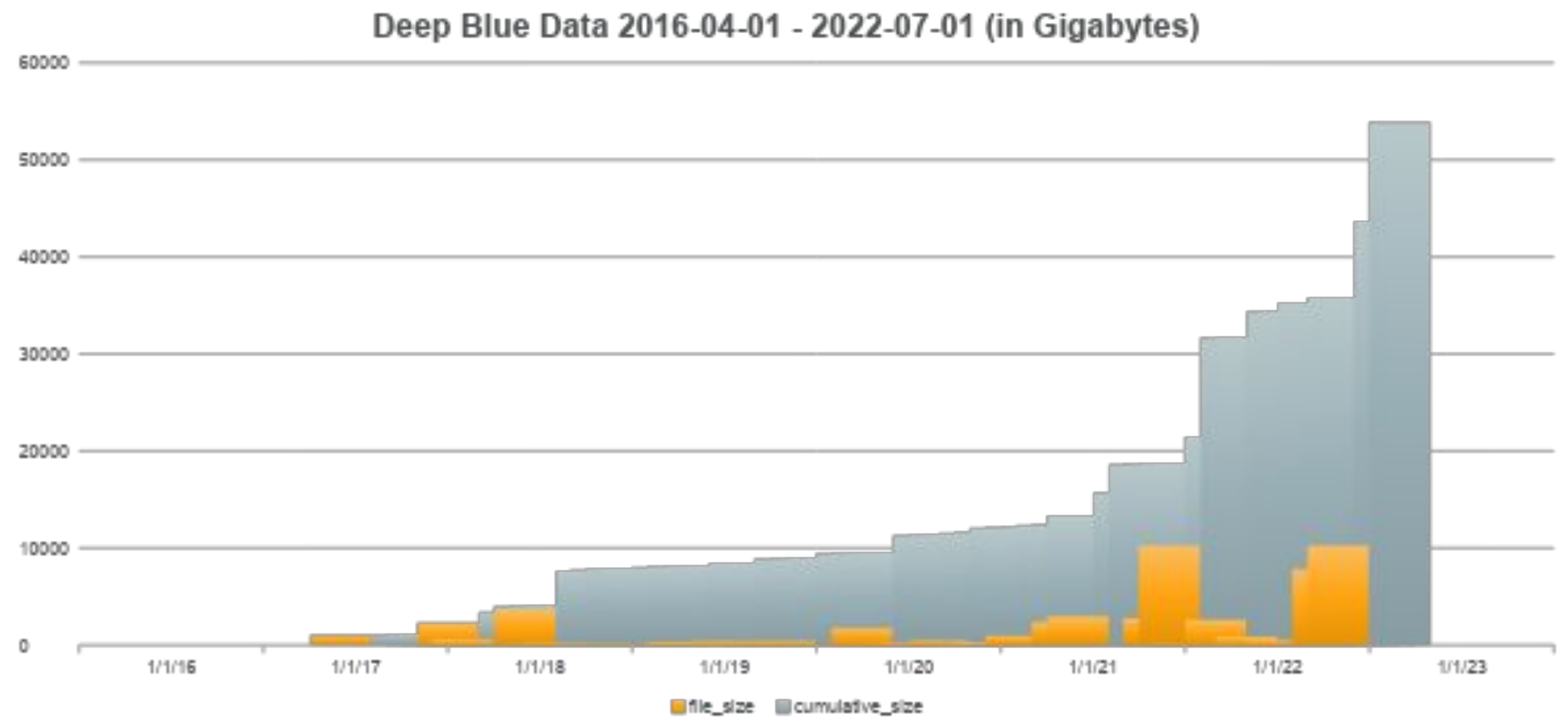
WHAT CURATORS NEED FROM RESEARCHERS

- Talk to us earlier in the lifecycle!
- Any kind of documentation at all!
- Ideally clear documentation
  - In the deposit metadata
  - In the README.txt

This is when we hear from researchers



https://data.research.cornell.edu/life-sciences


Deep Blue Data 2016-04-01 - 2022-07-01 (in Gigabytes)
60000
50000
40000
30000
20000
10000
0
1/1/16
1/1/17
1/1/18
1/1/19
1/1/20
1/1/21
1/1/22
1/1/23
file_size
cumulative_size

DATA CURATION NETWORK

https://datacurationnetwork.org/about/our-mission/

University of Minnesota
Duke University
PennState
Cornell University
VT University Libraries Virginia Tech
University of Nebraska Lincoln
The Michael J. Fox Foundation for Parkinson's Research
University of Michigan
Johns Hopkins University
Dryad
Illinois
NYU
UC Santa Barbara
Princeton University
Washington University in St. Louis University Libraries

Most of these institutions have 1- 3 people (FTE) doing data/software curation for their entire institutions those that have more are most likely expanding into library subject specialists who are doing it part-time. *Dryad is not tied to an institution so has more curators available.

RESOURCES I HAVE MADE AND SHARED

- GitHub repo with tips for working with large data sets: https://github.com/mutanthumb/LargeDataProcessing
- netCDF data curation format profile: https://deepblue.lib.umich.edu/handle/2027.42/145724
- MATLAB data curation format profile: https://deepblue.lib.umich.edu/handle/2027.42/154686
- Data sharing workflow for large datasets: https://deepblue.lib.umich.edu/handle/2027.42/142389

RESOURCES USED IN SLIDES

- Andrea K. Thomer, Dharma Akmon, Jeremy York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The craft and coordination of data curation: complicating "workflow" views of data science. 1, 1 (February 2022), 27 pages. https://arxiv.org/abs/2202.04560v1
- Acker, A, Donaldson, DR, Kriesberg, A, Thomer, A, Weber, N. Integrating research and teaching for data curation in iSchools. *Proc Assoc Inf Sci Technol*. 2020; 57:e285. https://doi.org/10.1002/pra2.285
- Labou, S., Otsuji, R., & Minor, D. (2021). UC San Diego Ithaka S+R Research Study: Supporting Big Data Research. *UC San Diego: Library*. Retrieved from https://escholarship.org/uc/item/8kr7p2c0
- Mullendore GL, Mayernik MS and Schuster DC (2021) Open Science Expectations for Simulation-Based Research. *Front. Clim.* 3:763420. doi: 10.3389/fclim.2021.763420

QUESTIONS?

Susan Borda

Digital Preservation Projects Manager

sborda@umich.edu

@mutanthumb