EarthCube Model Data RCN July 25-27, 2022 Workshop

07/25/22:

-Varying publisher expectations, how do we figure this out?
-EGU publishers are very stringent on what repositories can be used.
-Reviews, funding agencies, etc need guidance.
-What components are considered curation?

Plenary presentations, see:
https://modeldatarcn.github.io/workshop3/presentations/presentations.html
-Curation -Access to storage is an issue. No server access to manipulate data.
Need tools for repositories and curations "plug and play" software solutions for data curation.
-Documentation lacking
-Sharing workflows advances scientific integrity and the advancement of scientific knowledge. Research transparency. Examples of different results for the same question from the same datasets.
-Focus on preserving data from projects are important for downstream reuse. More important to share workflows.

-See Cindy B's slide on time saved when workflows were reused.
-The goal is science, not to store data.
-Data may not be the problem regarding reuse or building on findings, accessibility of workflows and code might be more important in some cases.

-Q: Is it important to store older versions of data that have been released in multiple versions, e.g. CMIP3/4/5/6? Older versions may be much smaller, being dwarfed by newer versions. Some people may have specialized code that use old data, because they don't have funding to upgrade their models to new versions. CMIP data sets may be outliers in that there is still community use of older versions, and comparison between versions. This may not be the case for other model data, or reanalyses, do people use old versions of reanalyses?

-People may not want to reproduce your science, they may be more likely to want to use your science to do something else. This involved learning from workflows and reusing workflows, less about learning from data or reusing data.

Q from Cindy's talk - while other scientists may be most interested in workflows and redoing work for another time/region/etc. but, are we missing a community of decision makers who want the output data in easily usable formats to support their work? E.g. a state climate adaptation officer who wants a well described/vetted dataset to plan for future climates.

-On requirements, what are the financial implications of journal requirements on PIs? If people don't have resources to meet the requirements, what are they supposed to do?

How do we get cross agency Pangeo like resources work with security requirements and the Anti-defficiency Act?

Breakout 1 -Sustainable Curation full notes

Breakout 1, Sustainable Curation Summary:

- **Software and data management plans need to be well thought out by PIs/creators and elevated in importance by funding agencies (broader impact).**
  - Why are we doing this?
  - Who to ask?
  - Once I know what data to keep, what should I ask myself?
- **Need to focus on making students and early career of importance of workflow & data preservation.**
  - And provide training, resources, examples, and support.
  - Added to curriculum?
- **Funding should come from agencies specifically for data/software management needs**
  - Data/software management in proposals.
  - Rubric can provide standardized guidance to limit subjectiveness from funders/proposal reviewers
- **Better discoverability - search engines (think google scholar for data / workflows)**
- **Better reusability**
  - support for Metadata, and thorough documentation

07/26/22:

Breakout 2 -Determining Lifetime for Simulation Data full notes

Breakout 2, Determining Lifetime for Simulation Data Summary:

- **Plan and advertise de-accession strategy at the point when data is deposited**
  - E.g., Metadata states, Data will be accessible for N years, upon which a de-accession decision will be made
- **Defined process to evaluate data value**
  - Examine direct and indirect metrics (data access and citation metrics)
  - Evaluate cost to continue to preserve data at the existing service level
  - Notify community that data will be de-accessed and turn off access to the data and see if the community complains
    - Engage community if there feedback, including pointing users to updated data products.
  - Move forward to purge the data or preserve on cheaper cold storage

    - Allow for PI to continue to support data archival and access
- **Incentivize researchers to move to newer versions of a data product**
    - Update legacy software dependencies on old data structures and formats
- **A rubric to evaluate whether or not data should be preserved moving forward**
- **Does the data fall into the category of knowledge production vs data production?**

Breakout 3 - Incentivising Data and Software Preservation and Sharing full notes

Breakout 3, Incentivising Data and Software Preservation and Sharing Summary:

- **Better communicate how much more successful researchers will be if they participate in Open Science**
    - Provide narratives of individual success stories
- **Update promotion and tenure process -need credit for doing open science**
    - What concrete metrics would support this? Demonstrate that “impactful” science includes more than just papers.
- **Develop more diverse/representative compositions on editorial boards, proposal review panels, etc.**
- **Helping researchers find the optimal repository**
- **Normalize the practice of code and data review. Normalize that this is a learning process for everyone.**
- **Raising the visibility of open science achievements**
    - Data/software related awards by societies
    - Data/software related conference themes
- **Collaborations and partnerships to achieve open science**

Breakout 4 -Equitable access to data and software curation and analysis resources full notes

Breakout 4, Equitable access to data and software curation and analysis resources summary:

- **Help people to comply to meet minimum publisher standards, instead of giving waivers.**
- **Accessible compute resources (data/bandwidth proximate)**
- **Accessible expertise / workforce development in scientific workflows and curation**
    - Train more people, and bring in people who already have this expertise
- **Community guidance on equity-focused practices**
- **Make workshops and training more accessible to people in underserved communities**
    - Bringing people to workshops, and/or sending workshops to more communities
- **“National virtual data curation laboratory”**
- **Better advertise what is out there now**
- **Invest in building relationships**
- **Well resourced Flagship Universities/Institutions to lead efforts in providing resources to smaller institutions**
    - Smaller/MSI institutions need to be empowered as equal partners

General Discussion:
Repository recommendation help desk supported by a coalition of publishers. Similar to NCAR dataset submission system. - See Data Help Desk - Collaboration between ESIP, AGU, EGU, AMS, ESA

Possible page to summarize summarize and allow contributions of community resources: https://github.com/modeldatarcn/modeldatarcn.github.io/blob/master/communityresources/resources.md

07/27/22:

9:00 am summary and next steps notes

National Virtual Data Curation Laboratory, similar idea to National Virtual Biotechnology Lab -https://science.osti.gov/nvbl

-What repositories does the community actually use?

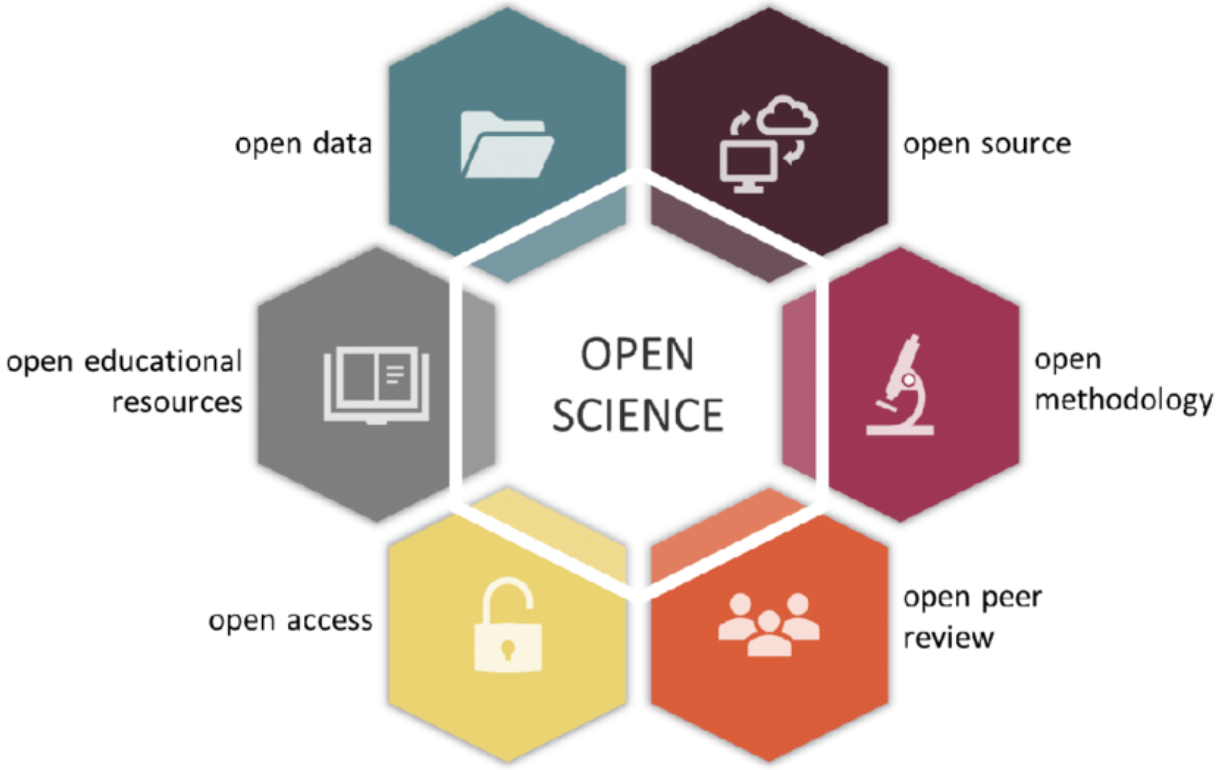
-AMS annual meeting -software and data preservation access statement fall all presenters?

Publisher policies, Agency DMP requirements -these should be consistent

What subject experts do we need to make the open science spiral forward work?:
Scientists, data curations expertise, software development/curation expertise, open science experts, educators, stakeholders who use actionable science
Open Science Support Network -Scott Collis

Need senior level science champions (PIs who have historically gotten large funding $) for open science to nudge and convince colleagues along. Nurture this group, and amplify the voices of those who already exist.

-



Open Science Spiral - life cycles that move forward in time. Avoiding gaps between revolutions of the spiral

Next steps
-follow on actionable workshops.
-decision makers on promotion evaluations

Open source science hackathon -open science tool development
-develop WRF modeling templates
Discussion on what people are doing now in the geosciences open science space?
General question -how well are we doing now?

Comments from early career and graduate student group
-PI don't appreciate the time spent cleaning up code.
-Need senior level champion in addition to early career grassroots groundswell.
-Demonstrate how reuse of well documented code and workflows is infinitely easier

-Data and software references required in proposals -past performance contribution to open science.

Mentoring is needed.

-Publisher policy review hackathon
--publishers engage their communities/stakeholders for feedback
-Research Data Alliance has discussions related to this topic. -somewhat closed
-Sponsor Data Management plan review hackathon

Tension between the message that these simulation outputs are important, but keeping all of the data forever is not possible/necessary.

-Consider what will need curation during the project planning phase.

-How can we support a “production, easier to use, version” of the rubric that can be used by publishers? Could a publisher, or coalition of publishers take this on?n

Wanted to plug our AGU session that might be of interest to everyone here: **ED024 - Open Science Practices and Success Stories Across the Earth, Space and Environmental Sciences**
https://agu.confex.com/agu/fm22/prelim.cgi/Session/161155

I'm hoping that one deliverable from this workshop might lead to clearer guidance from NSF with regard to writing an effective, "modern" Data Management Plan.

Kevin Tyle -Class to teach students about the open source ecosystem in the Atmospheric Sciences community. Get the next gen of scientists using open source tools for their science and contributing to development of those tools.

Teach and engage senior level generation of scientists about open science tooling, open source opportunities. They might not even realize these opportunities exist.

Clark E. -AMS Mesoscale processes conference in Madison July 2023. Embed in this conference to promote exposure -NASA tops.

See: AMS NSF “Mind the Gap” effort

Mention of registered reports in this community 300 journals participating.
https://www.cos.io/initiatives/registered-reports
-Tries to remove fear of being scooped:
-Publish work well before the final result
-Happens early in a project
-Publish science process that don’t produce expected results -science that doesn’t show anything

Label “open scientists” and “closed scientists” vs young and old scientists.

Use “open study design” vs “open science”

Hi everyone here is a pitch for checking if your organization/institution has signed on to DORA - The Declaration for Research Assessment which starts speaking to the Tenure and Promotion conversations https://sfdora.org/read/ here is the signatory list https://sfdora.org/signers/ see if you are there!

Pressure to publish at early career stage -carve out your niche to survive drives pressure against open science practices

Rubric

- Have a web survey version
- Weighting - have a way to decide relative weightings at the beginning when filling it out?
- Create an output that records the responses, could be compiled into a readme
- Who is filling it out?
    - Scientists individual vs Together with a data curator
    - Imposter syndrome vs center of the universe syndrome
- Point to an expert who can help you with the rubric if you aren’t sure how to answer, or send the rubric outcome to see if it makes sense. Would make it more likely that people would use it.
- NSF has guidance on how to do data management plans, maybe AGS could add this to their DMP guidance.
    - They send out a monthly email with announcements. Could this be sent out as an announcement?
    - Request people to try it out, provide feedback through some mechanism.
- A prototype web survey version of the rubric was developed by Fernando Rios at https://zoidy.shinyapps.io/ModelDataRubric/.

Thinking of the rubric, Minnesota has a Data Storage tool that helps folks work through their thought process of what storage to use for a specific project, which is what the rubric is doing more or less. A format like this seems user friendly and if the tool provided an output, that could be used as part of a data statement.

Emphasize documentation in Rubric usage instructions and reference use case testing template. Documentation that describes the comprehensive simulation experiment workflow and resultant data.

Video tutorial showing how to use the rubric and logic behind the rubric
Need to examine long term usability of “production” rubric product.
Example Rubric as a survey (R based app) by Fernando Rios:

Since it was brought up in the workshop and I had a few hours to kill at the airport, I coded up an interactive version of the rubric https://zoidy.shinyapps.io/ModelDataRubric/. I'm pulling the text straight from the rubric. This revealed that many items need to be reworded so that they are more appropriate for a questionnaire (i.e., standardize all of the descriptor definitions to be in the form of a question)