

**Chat Discussion** from the  
COPDESS Workshop: Rubric for Models and Model Data — Best Practices for Preservation and  
Replicability

10:05:52 From Shelley Stall to Everyone : Welcome everyone and thank you for participating.

10:16:44 From Shelley Stall to Everyone : The Steering Committee is available here:  
<https://modeldatarcn.github.io>

10:18:05 From Gretchen Mullendore to Everyone : Thanks, Shelley!

10:23:21 From Maggie Hellström to Everyone : In connection to "knowledge production", have you at all thought about methods & means to support annotation of model output datasets? This would allow users of model outputs and/or modelling-based publications to enrich the related (catalogue) metadata with comments on quality, applicability, boundaries/restrictions etc.? This could facilitate the capture of knowledge that is based on data and information - something that IMHO may be even more important for simulation/modelling than for observations. I'm thinking that publishers could play a very important role here, by providing platforms for (moderated or peer-reviewed) annotation & commenting of data in direct connection to articles.

10:29:38 From Chris Erdmann to Everyone : For those interested, since workflows has come up, GO FAIR US is running a FAIR Workflows session with Dockstore and WorkflowHub (in connection with the upcoming EarthCube Mtg) <https://gofair.us/events/2021-06-16-FAIR-Workflows.html>

10:30:16 From Suzanne Pierce to Everyone : Already, this approach looks like it will be really useful - thank you for all the hard work of pulling it together and documenting it to this point.

10:31:11 From Matthew Harp to Everyone : +1

10:34:29 From Chris Erdmann to Everyone : Matt, I like that you are using replicability! +1

10:40:17 From Chris Erdmann to Everyone : Per Matt's comment about Jupyter Notebooks, not just AGU but others around the world have been contributing to guidance that addresses some of the challenges he mentioned <http://doi.org/10.5281/zenodo.4774441> Some additional feedback from colleagues from Pangeo and NASA will be added shortly in a new version.

10:41:15 From Maggie Hellström to Everyone : Thanks Chris for sharing that link!

10:41:32 From Susan Borda to Everyone : <https://mybinder.org/> is nice resource for Jupyter Notebooks to go along with a repository

10:49:08 From Maggie Hellström to Everyone : It seems to me that the Rubric must be taken into account already at the Data Management Planning stage, especially concerning the cost estimates. Currently most people only think about the resources required for storage and computation during the research project active phase. This makes it very difficult for e.g., university data centers and data stewards to appreciate the commitment that they need to put in in order to support researchers in data-intensive disciplines.

10:51:00 From Susan Borda to Everyone : And they only think of sharing their data when publication time comes around.

### Chat Discussion from the

COPDESS Workshop: Rubric for Models and Model Data — Best Practices for Preservation and Replicability

10:51:04 From Varsha Khodiyar to Everyone : My view is that the minimum for a research article should be sufficient data/context to support the claims being made in the article. For a data paper, there should be sufficient data/context to support reuse.

10:51:12 From Chris Erdmann to Everyone : Exactly Maggie! We had this challenge on an NIH project I was on. We had projects sign up only to realize that the resources required were substantially more and we had to have follow up conversations about refining their proposals. Here is the form at least for estimating those costs, sending a pre-proposal <https://biodatacatalyst.nhlbi.nih.gov/resources/cloud-credits>

10:53:22 From Chris Erdmann to Everyone : Stephen, the rubric/website is highlighted in our AGU guidance :)

10:56:29 From Maggie Hellström to Everyone : While total storage volume is an important aspect, IMHO the "structure" and "granularity" of the data, software etc is perhaps even more crucial when it comes to reproducibility as well as "fitness for use" of data for others. (And indeed also for repositories!)

10:58:07 From Chris Erdmann to Everyone : Gretchen, Doug, Matt, can you share the link to the Google doc you were walking through Gretchen?

10:59:26 From Chris Erdmann to Everyone : Stephen, we have your example if you are willing to share from JAMES?

11:01:16 From Matthew Harp to Everyone : +1 Doug also research teams may be willing to share but 'don't have the time to produce the meta information and documentation to really help others understand the documentation - in a recent example we indeed compromised with referencing their methodology doc/pubs for simulation data

11:01:29 From Gretchen Mullendore to Everyone : @Chris: [https://docs.google.com/document/d/1SS2v\\_AaioYNU2eJ\\_GwKX49Wiwh3CatiV4-IQ2eXEhGU/edit](https://docs.google.com/document/d/1SS2v_AaioYNU2eJ_GwKX49Wiwh3CatiV4-IQ2eXEhGU/edit)

11:03:50 From Gretchen Mullendore to Everyone : All: the doc I just shared doesn't exactly match what's on our website, as still work in progress

11:04:31 From Matthew Harp to Everyone : +1 Maggie :-)

11:07:05 From Chris Erdmann to Everyone : Stephen shared his example recently <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS001726#jame20966-sec-0068-title>

11:07:41 From Chris Erdmann to Everyone : As we are building our AGU data and sharing guidance for JAMES.

11:10:43 From Shelley Stall to Everyone : Laurent is AGU's Editor-in-Chief for JGR: Planets

11:11:27 From Gretchen Mullendore to Everyone : @Laurent: I've heard that a lot as well. Fear of misuse of code/models.

11:11:38 From Gretchen Mullendore to Everyone : Great point to raise!

## Chat Discussion from the

### COPDESS Workshop: Rubric for Models and Model Data — Best Practices for Preservation and Replicability

11:14:12 From Chris Erdmann to Everyone : We saw quite an uptick in publishing software in JAMES in GitHub/Zenodo and NCAR/UCAR in 2020.

11:14:39 From Maggie Hellström to Everyone : @Laurent: very important point about the fear of misuse etc. Perhaps one way to alleviate their fears would be to make better use of the possibility (in standard metadata schema like Dublin Core) to outline the intended usage of a digital object. Then, if someone does something "stupid" with the data or code, then the original authors can always refer back to these usage notes...

11:15:12 From Pauline Chauvet to Everyone : +1 Maggie

11:16:09 From Chris Erdmann to Everyone : I've been seeing the friction points with ECRs wanting to release code vs their lab.

11:17:44 From Maggie Hellström to Everyone : Do those senior researchers at least agree to share reasonably detailed "pseudo-code" and flow charts, or do they keep these as well vague or fuzzy in order to prevent "scooping"?

11:18:25 From Susan Borda to Everyone : I'm glad to have a framework and vocabulary to use when I work with researchers.

11:19:52 From Laurent Montesi to Everyone : @Maggie, the specific example I have in mind, the person actually published a book from which it's possible to reconstruct their code (a couple people did it). In others, they describe the strategy but not the implementation.

11:19:57 From Suzanne Pierce to Everyone : Hopefully everyone is familiar with Yolanda Gil's leadership on this topic. There are many training resources, papers, and services via the EarthCube Ontosoft project: <https://ontosoft.org>

11:20:20 From Suzanne Pierce to Everyone : It includes digital objects of research with both software and data product consideration

11:22:31 From Chris Erdmann to Everyone : To add what Shelley mentions, we settled on 1 TB based on some of the general repos. Caveat here is we are still looking at this.

11:22:54 From Maggie Hellström to Everyone : What about peer review? Are reviewers and/or editors being given access to code in order to check the quality of the research + simulation output evaluation?

11:23:21 From Stephen Griffies to Everyone : Yes, that is at least the ideal.

11:23:36 From Chris Erdmann to Everyone : Currently Maggie that is being done through HPC Centers, sharing services like Box or Dropbox...

11:24:07 From Pramod Thupaki - CIOOS to Everyone : Thank you for a great presentation and discussion ! Sorry I have to duck out to attend another meeting.

11:25:27 From Chris Erdmann to Everyone : BTW Maggie, at NIH it is done in buckets like at Google and Amazon.

11:27:14 From Gretchen Mullendore to Everyone : @Suzanne: thanks for sharing Yolanda's project

### **Chat Discussion** from the

COPDESS Workshop: Rubric for Models and Model Data — Best Practices for Preservation and Replicability

11:28:04 From Susan Borda to Everyone : I have a ~16TB dataset in our repository at the UMich Library and have in a tiered storage where the most critical part of it is “live” in our production system and the rest of the ~13TB is in a tape environment that is replicated only and a lot less expensive. Using Globus for download.

11:30:34 From Suzanne Pierce to Everyone : Nice Susan - we’ve got a project called DOLCE (Digital Object LifeCycle Ecosystem) that allows users to create Digital Collections that are anything >10GBs and store it on TACC resources then push the collection to Dataverse where the libraries can maintain it - we’ve agreed to 10 years minimum to maintain and store the objects/collections

11:30:49 From Maggie Hellström to Everyone : I will for sure use the rubric in my RDM teaching!

11:30:52 From Matthew Harp to Everyone : Thank you!

11:30:56 From Pauline Chauvet to Everyone : thanks !!!

11:31:03 From Suzanne Pierce to Everyone : Thank you this is wonderful - great discussion!