Repo Renovation: motivation, methods, and lessons learned in ‘doing the FAIR thing’

Steve Diggs

Thursday, August 22, 2019, 10:30 PST (17:30 UTC)

Connection Information

https://global.gotomeeting.com/join/157892821
You can also dial in using your phone.
United States: +1 (312) 757-3121
Access Code: 157-892-821

Presentation

Slides
Recording

Attendees

Jocelyn Elya, Carolina Berys-Gonzalez, Susan Becker, Andrew Barna, Debbie Roth, Steve Diggs, Suzanne O’Hara, Megan Carter, James H. Swift, Mathew Biddle, Joseph Gum, Chris Olson, Bob Key, Alex Kozyr

Minutes

1. FAIR = Findable Accessible Interoperable Reusable
    a. Today focusing on F and I
2. Over the course of the last 50 years, data volume has exploded and cost has decreased.

    a. Cost decrease allowed metadata attitude to change
    b. SC2 format. Station data metadata format example.
3. WOCE
    a. A series of in-person meetings. Version 3 of data that made data more interoperable. Converged around COARDS and netCDF
4. Repositories can
    a. Take inventory of what they have now, design new architecture, plan, switch from old to new, and maintain
    b. Timing is key
5. Findable - focusing on schema.org P418. Interoperable - focusing on CF.
6. CF convention
    a. Unidata is the custodian in the US
    b. Descendant of COARDS
    c. Most widely implemented in netCDF
7. WHP-exchange format (2001)
    a. Metadata were scattered around
8. Then to COARDS-compatible netCDF
9. A conversion to these standards - Is it worth it? And how much is it going to cost me?
    a. Scope it properly
    b. Decide on a strategic and desirable “why” it’s needed
    c. Ask for the resources you need.
        i. ~50K - $125K, 18 - 24 months
    d. Look for “broader impacts”
10. Temptation is to swing for the fences and get it done in one shot. But allow for incremental improvements.

Questions/Discussion

1. Bob Key - Crunching the budget numbers. Are these realistic?
    a. Supplement existing projects with this.
    b. Can you even “turn the crank” for that much money? Need to know which files need reprocessing.
    c. Andrew Barna says they have ~15,000 files
    d. Incremental and slow. Can’t do all data at once.

    e. “¾ of what we need to do we can do off existing resources. And the rest is witchcraft”
2. Europe has got pangea trying to undergo renovations. Mathew Biddle BCO-DMO has been migrating to an ERDDAP server.
3. WOCE Version 3 was expensive to achieve. Flying people around the world to converge on technical solutions.
4. How do you know what the research community wants?
5. Bob - do you make the data offerings yourself, or the tools that someone can use to create them?
    a. Talking about using ERDDAP with netCDF file to create different files.
6. Headache for users is reading the code in. I/O, not analysis
    a. Going forward, supply software with each dataset. Python and MATLAB at least.
7. Jocelyn - Any plans for formalize how to do this?
    a. Yes, a how-to with EarthCube
8. Mathew Biddle - at BCO-DMO they’re migrating data servers
    a. Migrating to ERDDAP
    b. Want to use federated search features
9. A big part of the budget has to be maintenance
10. Chris Olson - doing this transformation in parallel with operational tasks. No defined start and end. How did that work?
    a. Some context switching. Sometimes completely devoted to routine operations, sometimes completely devoted to innovation. Devote enough time to make some progress in each.
    b. Split the team in terms of focus. But no one person does one thing.
        i. Need more eyes on the problem. Better than having only one person looking.
11.