Council of Data Facilities
4th Wednesday of the month
11:00 PM Pacific / 2:00 PM Eastern

Mailing List: https://lists.esipfed.org/mailman/listinfo/esip-cdf (#council-of-data-facilities)
Slack: https://esip-all.slack.com/messages/C8EACC8KT

Zoom Meeting:
Get the current meeting link from the ESIP community calendar:
https://www.esipfed.org/community-xcalendar/

Presentation Recordings:
https://docs.google.com/document/d/1RmSvZ5DLmkjLN-Rm_jjWQoD_YPwigC2cpENqoNxVzHY
Future presentations:
https://docs.google.com/spreadsheets/d/1tBwsuEdl-h2Wy0b_N-ORdfwPXCMpWLFs-om0Hid1HQw

Sep 24, 2025 2:00 PM EST

Attendees:
Nick Jarboe, EarthRef/MagIC
Rob Casey, EarthScope
Martin Seul,
Matt Jones, DataOne
Karen Stocks
Gregory Maurer
Aaron Friesz

Agenda: Shared resource discussion.

Notes:

OCE Ocean Science. has office hours on the 30th Sept.

Interest in having a talk on slowing down AI scrapers.

https://developers.cloudflare.com/bots/concepts/bot/#ai-bots

https://www.nsf.gov/funding/opportunities/pcl-test-bed-test-bed-toward-network-programmable-cloud-laboratories?sf228067724=1
NSF PCL Test Bed looking for proposals. Looks like AI is a component of this.

https://ci-compass.org/services/

Science gateways NSF

https://blog.barracuda.com/2024/11/19/threat-spotlight-bad-bots-evolving-more-human

Curating the data to have a high quality subset of datasets for AI

AI-Ready Data overview for researchers:
https://cyber2a.github.io/cyber2a-course/sections/ai-ready-data-in-arctic-research.html

And the ESIP AI-readiness checklist: https://doi.org/10.6084/m9.figshare.19983722.v1

Open storage network OSN/OSDF- NorthEast Storage exchange - Harvard and others
DataPreservaiton network -
List of archive sites - Dark archives

Clocks group - dark archive for journals
VPN - for datasets -

Karen - I mis-spoke earlier, the NESE tape group is not part of OSN. They are part of some other research data storage consortium that I can't remember at the moment…https://nese.mghpcc.org/

Open storage network - buy in for a full pod - 1.4 Petabyte

We have a Funding Friday Project that seems relevant (maybe not) to this discussion
https://drive.google.com/file/d/1nJRe1eV-DsrjeXoHIWQC3yWpU5mvQn6e/view?usp=sharing

OSN and
OSDF

OSDF connected to pelican

Have a citizen science network for data storage and retrieval.
Open data space - Academic dataset

I gave an ESIP talk on content-based Identifiers a while back that touched on these issues:
https://zenodo.org/records/8166522

Aug 27, 2025 2:00 PM EST

Attendees: Nick Jarboe

Vishu Nandigam, UC San Diego
Shelley Stall, AGU
Greg Maurer, New Mexico State Univ. EDI Repository
Colin Smith, University of Wisconsin-Madison. EDI Repository
Karen S tocks, Scripps Inst. of Oceanography (for R2R and CCHDO)
Tom Cram, NSF NCAR
Mark Parsons, ESIP
Aaron Friesz, ESIP
Frank Nitsche, Columbia University (USAP-DC and R2R)
Rob Casey
Kerstin Lehnert
Clair Stirm

Agenda: Updates from Shelley Stall. Presentation of the shared resource survey results and discussion.

ESIP Announcements:
- ESIP July Meeting Recording - https://youtube.com/playlist?list=PL8X9E6I5_i8gkfxDKDXMIJNw7L1GbjmiT&feature=shared
- ESIP is seeking an Executive Director! (https://www.esipfed.org/executive-director/)

Notes: People outside the US concerned about data set losses.

Start consortium with donors but need to be funded by products in the long run
Five day convening to figure out what can/should be done.
Two parts - NSF funded repositories -
Adam Shepherd - share this survey
How to package the steps for multiple funding providers
Mark Parsons
Science-on-schema.org
Polar data search (https://search.polder.info/) - using science-on-schema.org

First responders, fire fighting,
This discussion reminds me of some presentation/discussions with Jed Sundwall (Radiant Earth) around data cooperatives. He spoke a bit on this during his Plenary talk at the summer meeting where he plugged Source Cooperative (https://radiant.earth/blog/2023/10/what-is-source-cooperative/). He might be someone to bring into the discussion.

AGU Impactful Datasets submission form:
https://docs.google.com/forms/d/e/1FAIpQLSf0Tlb1a29C2I-INyl8qf_1oG1tFBosYW5qGBBj70D42Nhn_w/viewform

Survey Results: Here is the survey https://docs.google.com/spreadsheets/d/1Tv6mQpkaS3NuELL9GFFVXR3Mkb4u00-KOpDbpgK6lG8/edit?resourcekey=&gid=1990474970

Jun 25, 2025 11:00 AM PDT

Attendees: Nick Jarboe, Karen Stocks, Martin Seul

Notes:
Do a round table intro
Somehow get some shared infrastructure to support

Dataone $13k per year.
One time import fee

European systems to support
Sloan Foundation
Gordan Getty foundation
Moore Foundation
What are the largest costs
What to share
Deep storage
OSN
Jetstream
Cloudbank
OSDF

Peter McCartney - Repository Sustainability

Merge #2 and #3 and bring up the other session that are dealing with these issues.
Look at Matt Jones session

Invite some repisentives

Look at which session are dealing with the

May 28, 2025 11:00 AM PDT

Attendees: Nick Jarboe, Karen Stocks, Rukaya Sarah Johaadien, Carolina Berys-Gonzalez, Maria Esteva

Presentations:
Rukaya Sarah Johaadien (Natural History Museum, U. of Oslo) - “ChatIPT tool to transform spreadsheets into standardized GBIF datasets”

Carolina Berys-Gonzalez (UCSD/SIO/CCHDO) - “NLP tool CCHDO developed to extract structured metadata from text cruise reports”

Maria Esteva (Texas Advanced Computing Center, UT Austin) - “A Data Curation, Interrogation, and Access System for the Texas Robotics DataVerse”

Notes: Update ESIP website with new time of the monthly meeting

Apr 23, 2025 11:00 AM PDT

Attendees: Nick Jarboe, Joseph Gum, Karen Stocks, Rob Casey, Matt Jones, Gregory Maurer

Birds of a feather meeting at ESIP - About CDF meeting -

Co-run Scorecard session - Sustainable database cluster
Session on Scorecard and database survival plans.
CDF stand alone
ESIP general for NOAA decommission list

Matt-FYI, several of us are putting together an automating data quality session as well that might be of relevance

Deadline - May 15th

Discussion Topics:

1) 10 minute presentation and then discussion on Repository Crisis Scorecards (RCS) by Joseph Bum (NCAR).

The Repository Crisis Scorecards (RCS) are meant to measure how resilient a repository might be in its normal state and during certain crises. This includes a measure of how well a repository

might weather an example crisis, how easy it might be to restore metadata, and how much societal impact a missing repository would have. The scorecards are based off of the model data preservation rubric developed by Schuster et al, 2023

2) What session, if any, should the CDF run at this summer's ESIP meeting?

Feb 19, 2025 12:00 PM PST

Attendees: Nick Jarboe, Karen Stocks, Matt Jones, Vishu Nandigam, Megan, Aaron.

Discussion Topic: Future CDF activities

- Possibly time is bad - but everyone is busy
- Interest in working together on a common goal, e.g. P418 (then moved to science on schema.org). Vishu would like to propose a similar initiative. He suggests Croissant metadata. Follow up with effort to roll out across data repos. A common mission brings people together. Not clear who would lead.
- Some sense of weariness in the SOSO team (and other ESIP clusters)
- Small core funds to lead initiatives can help keep momentum. ESIP funding friday or lab grant awards are small pots of money.
- Does Google (Data Commons; google.org has climate impacts program) have grant programs that fit? Other commercial companies that may want "AI ready data". Two 50% FTEs would help.
- White paper on, e.g., top 10 interoperability priorities for data repositories, what work is being done, etc. (some white papers came out of CDF in the past - look at work to build)
- Any interest on ways to demonstrate the value/impact of data repositories? Learning from each other on what metrics, etc. (ESIP, AGU sessions have covered this, including tools, etc.). Matt has some automated tools for searching lit (see https://dataoneorg.r-universe.dev/scythe) - gets a lot more than the publisher reporting. Also the value of specific science stories vs just stats.
    - See also COUNTER Code of Practice for Research Data for standard metrics on data access
    - Best practices to coordinate metrics on impact of repos…if everyone is counting different, they aren't comparable.
- Strength of CDF is identifying areas where interoperability is important, moving that forward as a community.
- Goal: have substantial in-person CDF mtg at summer ESIP, we can use some meetings between now and then to plan it.

Topic Ideas

- Cloud migration. Whether to make the leap, what to watch for, how to organize staff, sustainability and uptime.
- Data Publication/Citation: Take a proposal to the publishers that data repositories can actually implement. Researchers should be able to publish a recipe (or reliquary) of how they accessed the data, formatted such that the recipe itself could be a standardized request format. [Repositories are not going to store researcher derivative data sets for open access, nor do we want everyone encouraged to run to Zenodo, which stores unstructured black boxes.]
- Utilization of AI/ML in ways that help 'repositories' function better. For instance, could we develop processes that generate domain-expert ML models to assist with semantic search and QA functions?
- Edge Computing in service of managing remote instrumentation and promoting standardized data collection.

January 15th

Attendees: Nick Jarboe, Reyna Jenkyns, Chenyue Jiao, Anna Kelbert, Edwin Henneken, Frank Nitsche, Greg Maurer, Hannah Blanco, Jess Tate, Karen Stocks, Leslie Hsu, Lesley Wyborn, Danie Kinkade, Robert Downs

Meeting presentation:
Anna Kelbert from the NASA Astrophysics Data System (https://scixplorer.org/)
anna.kelbert@cfa.harvard.edu

Title: Science Explorer (SciX): A Valuable New Information Service for Earth Science Literature, Data, and Software Discovery

Abstract:
This presentation will introduce the Science Explorer (SciX), a newly expanded, NASA-funded open science resource designed to expose and integrate scientific literature from all domains of the NASA Science Mission Directorate within a semantically-enabled platform, and to enhance this information with linkages to data and software. Building on the strengths of the Astrophysics Data System, SciX now encompasses Astrophysics, Planetary Science, Heliophysics, Earth Science, and the Biological & Physical Sciences. The platform features robust reference management tools and advanced bibliometric data visualizations intended to aid researchers in uncovering new research connections.

Next Month:

Discuss the future of the CDF.

Draft email to send to the CDF members:

Hello CDF,

As we look to the new year, we would like to have a CDF call where we talk as a community about our priorities and interests in the CDF. For the last couple of years, we have worked to schedule webinars from repositories or from related programs and activities. Overall, attendance has been reducing. Let's brainstorm about some options for CDF activities in the future at our [decide on month] CDF call.

We want to hear from you, but to get the thinking started, here are some options for future CDF meeting foci:

-White papers/knowledge summaries: identify a topic where we either want to speak as a community to outside groups (e.g. to funders, to publishers), or we think it would be valuable to bring together our knowledge on a topic to help data facilities. An example was the Registration Considerations document we produced in 2023. This could be something that rolls out over several meetings and/or is done within a working group. We could generate the topics ourselves, and we could also ask groups like NSF if they have topics they are interested in.

- Panels on topics of common interest, with multiple brief speakers and extended discussion time. E.g. "How is AI being used by repositories right now"?

- Or a similarly, "Tool Time": multiple short presentations from members where each presents on a different tool (approach, standard, technique, etc.), and how it has, or hasn't, helped that data facility.

What would make CDF something you are excited about attending?

November 20th

Attendees: Meredith Goins, Nick Jarboe, Joan Damerow, Chenyue Jiao, Cian Dawson, Rachael Blake, Daniel Segessenman

Meeting presentation:
Joan Damerow from the Lawrence Berkeley National Lab will present "ESS-DIVE: Community-Centered Data Repository for Interdisciplinary, Environmental System Science".

Abstract:

The Environmental System Science – Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) repository stores data supported by the U.S. Department of Energy (DOE) Environmental System Science (ESS) program. The ESS-DIVE mission is to preserve, expand access to, and improve usability of data generated through this program to address some of society's most

pressing energy and environmental challenges. Given the breadth and volumes of DOE research data, ESS-DIVE addresses the increasing demand for robust data management systems that can handle complex environmental data spanning diverse, heterogenous formats and large volumes. ESS-DIVE stores data spanning a variety of natural and human ecosystems such as watersheds, coastal systems and cities. Leveraging our standard reporting formats, we developed a novel Fusion Database – a service that extracts, indexes, and serves data from within standardized files, to enable advanced search beyond metadata. Future development of automated tools to help data contributors standardize their data and metadata will continue to lower the barriers to generating well-curated, rapidly reusable, FAIR data. We are also collaborating across five data systems supporting DOE's Biological and Environmental Program to enhance search and discovery of related environmental, multi-omics, molecular, atmospheric data, and more to support interdisciplinary science use cases.

October 16th

Attendees: Chenyue Jiao, Matt Mayernik, Daniel Segessenman, Nick Jarboe, John Paden, Martin Seul, Matt Jones, Natalie Raia, Tian, Walterj, Yanique Campbell, Danie Kinkade, Maria Esteva, Frank Nitsche

Agenda:

- Any other data repositories interested for dates in the new year.
- We hope to write some white papers on topics of interest in the coming year.
- Talk next month by Joan Damerow from Lawrence Berkeley National Lab on ESS-DIVE.

Meeting presentation:
Matthew Mayernik from the NSF National Center for Atmospheric Research will present “Enabling FAIR Facilities and Instruments via Persistent Identifiers”

- Presentation slides
- Project website

Persistent Identifiers (PIDs) are central to the vision of open science described in the FAIR Principles. However, the use of PIDs for scientific instruments and facilities is decentralized and fragmented. This talk will discuss a NSF-funded project to develop community-based standards, guidelines, and best practices for how and why PIDs can be assigned to facilities and instruments. We will outline findings in four main areas: developing a better understanding of the current PID ecosystem; clarifying how and when PIDs could be assigned to scientific instruments and facilities; challenges and barriers involved with assigning PIDs; and incentives for researchers, facility managers, and other stakeholders to encourage the use of PIDs.

September 18th

Attendees: Nick Jarboe, Chenyue Jiao, Danie Kinkade, Karen Stocks, Gregory Maurer, Melissa Cragin, Natalie Raia, Rachael Blake, Matt Mayernik, Bob Downs, Daniel Segessenman, Joan Damerow, John Beck, Tamar Norkin, Douglas Fil, Jerry Carter, Rob Casey

Agenda:

Future CDF work products. Should the CDF produce white papers on current topics of interest?
Example: **User Registration and Authentication**
https://docs.google.com/document/d/1YrwHUlBtJEO0Qju4Q8a8Blk7065QGDcJ854u0Zt8-5s

Meeting presentation:

Danie Kinkade from Woods Hole Oceanographic Institution will present her talk "Introduction to the RDA Repo2Pub Working Group".

Abstract: Increasing effort is being placed on streamlining research data publication, especially with respect to data that support scholarly publications. Yet, we still do not have a complete understanding of the highly variable data publication process, including the requirements and detailed workflows of individual stakeholders, such as the journal editor, author, and repository, and the touchpoints between them. Bottlenecks are arising at these overlapping points of stakeholder interaction, as domain and institutional repositories are becoming increasingly challenged to fit into the scholarly publication workflow. This presentation will introduce a newly formed Research Data Alliance (RDA) Working Group aimed at documenting critical touchpoints between data publication stakeholder workflows and their dependencies. This group plans to build on previous RDA work to provide much needed guidance for journal editors, repositories, and research authors to help streamline research data sharing in support of the scholarly publication process.

Notes:
Repo2Pub Working Group website:
https://www.rd-alliance.org/groups/coordinating-earth-space-and-environmental-science-data-preservation-and-scholarly-publication-processes-wg/

ESIP/RDA ESES Interest Group website:
https://www.rd-alliance.org/groups/esiprda-earth-space-and-environmental-sciences-ig/forum/topic/eses-ig-session-at-p19/

Create/log in to the RDA site and select "Join group" on these pages to receive updates about each group, respectively.

Our case statement/rationale outlines the full plan and proposed outputs Danie is outlining here and is downloadable at:
https://www.rd-alliance.org/rationale/coordinating-earth-space-and-environmental-science-data-preservation-and-scholarly-publication-processes-wg/

Draft of recommendations link:

Dryad positioned as place for domain data when a domain repository is not available.
https://datadryad.org/stash/join_us#our-membership

August 21st

Attendees: Nick Jarboe, Doug Fils, Lynne Schreiber, Chenyue Jiao, Jeffrey Glatstein, Karen Stocks, John Beck, Matt Jones, Martin Seul, Ethan Davis, Gregory Maurer, Danie Kinkade, Renya Jenkyns, Ginger, Ezra Cheruiyot

Agenda/Notes:

Future CDF work products. Should the CDF produce white papers on current topics of interest?
Example: **User Registration and Authentication**
https://docs.google.com/document/d/1YrwHUlBtJEO0Qju4Q8a8BIk7065QGDcJ854u0Zt8-5s

Meeting presentation:
Doug Fils - A Perspective on the Schema.org Ecosystem
https://docs.google.com/presentation/d/1kgYYoMZyQbLKNDhBzgN6g9otLkODoIiRK_wRVGX9u5o/edit?usp=sharing

A brief update on some scheme.org based profiles the author is aware of. These include updates on the ESIP Science on Schema.org work as well as developments in Croissant, CODATA Cross Domain Interoperability Framework (CDIF), the UNESCO ODIS Ocean InfoHub, POLDER, and other related projects. A perspective on some of the relations between these projects, or lack thereof, will be provided.

Slides:
https://docs.google.com/presentation/d/1kgYYoMZyQbLKNDhBzgN6g9otLkODoIiRK_wRVGX9u5o/edit?usp=sharing

DeCODER: https://www.earthcube.org/decoder

May 15th

Attendees: Nick Jarboe, John Beck, Bob Downs, Karen Stocks, Christine Kirkpatrick, Reyna Jenkyns, Lynne Schreiber, Dru Clark, Martin Seul, Joan Damerow, Alan Yang, Melissa Cragin, Joan Damerow, Rob Casey

Agenda/Notes:
ESIP Summer meeting session: Submission form info and notes
Title: Embracing Expanded DataCite Schema Support for IGSNs, Instruments, Publishers, and Funder PIDs
Session lead: Reyna Jenkyns
Other organizers: Nick Jarboe, TBD

Program:
Reyna Jenkyns, reyna@oceannetworks.ca - overview and moderation, publisher PIDs
TBD - Instrument PIDS
Nick Jarboe/other? - IGSNs using the new DataCite infrastructure
Kerstin/other IEDA person - transition from legacy IGSN to DataCite
TBD - Funder PIDs
Breakouts if necessary for discussion

Monthly Presentation:
Christine Kirkpatrick from the San Diego Supercomputer Center at the University of California San Diego will present on "Couch cushions, coupons & easter eggs: Ways to extend your research computing budget using NSF-funded (free) programs".

openstoragenetwork.org 10-50 Terabytes available

CloudBank - www.cloudbank.org - Can get up to 12k funds - overhead free
Contact: support@cloudbank.org
Naomi Alterman - naomila@uw.edu
Rob Fatland - rob5@uw.edu
Three page application

SGX3: Gateways to access supercomputing - They have services to support NSF work
National Science Data Fabric - Resource for creating software, data centric
FARR Workshop, OCT 9-10 - tinyurl.com/FARRinDC

April 17th

Attendees: Natalie Raia, Nick Jarboe, Karen Stocks, Reyna Jenkyns, Lynne Schreiber, Doug Fils, Megan Rush, Lindsay Powers, Hannah Blanco, Chenyuo Jiao, Martin Seul, Bob Downs, Tian, Greg Maurer

Agenda/Notes:

Karen Stocks - The FARR project is developing a webinar series around AI-readiness for data repositories. If your repository has taken steps to improve their AI-readiness, or has successfully created connections to AI users, or done other work in the area, and you might be willing to talk about it, please contact Karen at kstocks@ucsd.edu.

CDF have a small part of the GEO-OSE+ RCN Proposal. 3 student and 2 other travel grants to an annual meeting.

What type of ESIP session should the CDF host/support? Due May 10th
Data DOI introduction
Emphasize relationship metadata - parent and child relationships between PIDs, potentially also highlighting usage analysis being enabled by efforts through Global Citation Corpus
RDA working group on Data Repository Attributes and recs on metadata. (Re3Data and WDS planning to coordinate on adoption work and could help facilitate this idea)
DataCite would probably participate remotely Rorie Edmonds is the Samples Community Manager "rorie.edmunds@datacite.org".
Add more geo-specific
Sharing metadata and vocabularies

Natalie Raia from the University of Arizona, "Coordinating Data Preservation and Scholarly Publication Processes: A New Proposed RDA Working Group" (slides)

- WG draft case statement- please add suggestions via comments
- Join the RDA WG group to receive email updates about upcoming meetings & activities

Next case statement discussion: April 24th, 5-6pm ET
Please contact if you wish to share use cases, challenges, needs, etc. nraia@arizona.edu

March 20st

Attendees: CJ Woodford, Nick Jarboe, Reyna Jenkyns, Chenyue Jiao, Karen Stocks, Susan Shingledecker, Doug Schuster, Doug Fils, Frank Nitsche, Kerstin Lehnert, Matt Mayernik, Rebecca Koskela, Bob Downs, Tian,

Agenda/Notes:

CJ Woodford from the World Data System presents:
International Technology Office will present on the Global Open Research Commons Report Slides:
https://docs.google.com/presentation/d/1_ZuaP3evW2ax2HkkI-HRe4vd3NL54SeTTKBe4_eSVQE/edit

Feb 21st

Attendees: Douglas Rao, Karen Stocks, Nick Jarboe, Megan Carter, Jeffrey Glatstein, Alan Yang, Dru Clark, Frank Nitsche, Leslie Hsu, Matt Mayernik, Megan Rush, Melissa Cragin, Bob Downs, Lynne Schreiber, Chenyue Jiao, Martin Seul, Doug Schuster, Jon Weers, Yaxing Wei, Jessica Burnett, Reyna Jenkyns

Agenda/Notes:

Yuhan (Douglas) Rao at North Carolina Institute for Climate Studies will present on “What does "AI-readiness" mean for geoscience data repositories?”.

Future workshop links:
https://www2.cisl.ucar.edu/events/innovations-open-science-ios-planning-workshop-community-expectations-geoscience-data
https://www.farr-rcn.org/workshop24

Join the ESIP Meeting Highlights Webinar coming up in just an hour - fast-paced report-outs from nearly all of the recent January Meeting sessions. Find out how to get involved and which recordings to watch from the meeting (now found on our YouTube Channel). Check out the lineup and learn more at https://docs.google.com/document/d/1TvnZ4jvg4CXS0OUk6XFX1N8clPwvbgVbe1YO7hgPz3w/edit?usp=sharing.

Jan 17th

Attendees: Nick Jarboe (chair), Chenyue Jiao (Fellow), Robert Downs, Jerry Carter, Reyna Jenkyns, Geoffrey Stano, Edwin Henneken, Amber Budden, Martin Seul, Frank Nitsche, Karen Stocks, Megan Orlando, Matt Mayernik, Tian, Emilio Mayorga, Danie Kinkade

Agenda/Notes:

Introduction of Chenyue Jiao, a new ESIP community fellow that is joining the CDF.

Spreadsheet for future presentations. Would like to have CDF members volunteer to organize speakers for some future talks.

Bob Downs will be presenting:
“Managing Rights and Ethical Responsibilities for Sharing Open Data in the Earth Sciences”

Dec - canceled for AGU

Nov - canceled

Oct 18th

Attendees: Jerry Carter, Karen Stocks, Kerstin Lehnert, Lynne Schreiber, Megan Carter, Shelley Stall, Matt Jones, Frank Nitsche.

Agenda/Notes:

ESIP’s call for winter sessions is out. Should we submit one? Due Mon Nov 13.
Meeting start Jan 22nd.

ESIP session template:
https://docs.google.com/document/d/1F4IKJJlsOzBdJD2EhuimmnEV0Ej24XaulOPVcD8U2TE/edit?usp=sharing

Archiving Image data.

Need a speaker for November

Doug Rao from the FARR project has offered to present on the outcomes of the summer 2023 “Building Upon the EarthCube Community” Workshop session on “Working Towards AI-Ready Geoscience Data Repositories”, and continue a discussion of repository needs in this area.

Announcements

- GO FAIR US is offering AGU travel support awards - see application. Early career people, who would not otherwise be able to attend (and do not have an accepted abstract), with an interest in FAIR data, are particularly encouraged to apply.
- FARR is developing a short resource list around AI-readiness for repositories - input on the draft is welcomed, please send to Karen Stocks.
- Data pavilion presentations: any repositories interested in giving a lightning talk o AI-readiness experiences?

A presentation by Jerry Carter from the EarthScope Consortium titled "Dealing with Large Seismological Datasets: Clouds on the Horizon".

Christine Kirkpatrick will present in February:

“Couch cushions, coupons & easter eggs: Ways to extend your research computing budget using NSF-funded (free) programs”

Sept 20th

Attendees: Nick Jarboe, Shelley Stall, Frank Nitsche, Megan Carter, Kristina Vrouwenvelder, Christine Laney, Danie Kinkade, Doug Fils, Jeanette Clark, Karen Stocks, Lynne Schreiber, Martin Seul, Natalie Raia, Rob Casey, Bob Downs, Susan Shingledicker

Agenda/Notes:

Call for ESIP fellows. https://www.esipfed.org/get-involved/student-opportunities

Dear Colleague letter from GEO. $50k grants for data repositories.
AGU can write support letters for this grant.
https://www.nsf.gov/pubs/2023/nsf23141/nsf23141.jsp?WT.mc_ev=click&WT.mc_id=&utm_medium=email&utm_source=govdelivery

“Journal Production Guidance for Software and Data Citations”
Preprint
https://essopenarchive.org/users/536571/articles/616035-journal-production-guidance-for-software-and-data-citations

Open science pavilion - Reserve a Pod at AGU -

November possible talk - Joan
Nick - one suggestion for a future talk, possibly November, is to invite the ESIP Physical Samples Cluster to talk about the Sample Citation Guidelines and publication they are working on.

Shelly Stall presents:
"When is a data citation NOT a data citation: the good, the bad, and the machine-actionable"

August 16th

Attendees: Nick Jarboe, Karen Stocks, Shelley Stall, Lynne Schreiber, Jerry Carter

Agenda/Notes:

Shelly Stall - Will give a talk next month, Sept 20th, about about DataCite and how journals and data repositories reference each other and workflows.

NASA - Data centers don't want to store sub-datasets that people use for research papers. They are just going into general data repositories.

Jerry Carter - 18th of October - Talk on large datasets in the commercial cloud

June 21

Attendees: Meredith Goins, Nick Jarboe, Christine Laney, Jerry Carter, Karen Stocks, Matt Mayernik, Mike Bobak, Susan Shingledecker, Chris Jenkins, Megan Carter, Lynne Schreiber, Bob Downs, Chris Crosby, Corinna Gries, Shelley Stall

Agenda/Notes:
Meredith Goins from WDS (World Data System) will present on their recent review of strategic plans and technical roadmaps of WDS members, which had the goal of identifying needs and challenges, as well as testing their organizational assessment methods.

Presentation slides:
https://drive.google.com/drive/folders/1Pd5da-NXETLt9ASJmPBGZ8rbpiPjtNz1

Our session is at **4:00 pm on Tuesday: https://sched.co/1NocO**
Review of and planning for the CDF session at ESIP.

NSF public access plan 2.0 recently released: https://new.nsf.gov/public-access

Next month:
ESIP summer meeting.

May 17

Attendees: Bob Downs, Megan Carter, Shelley Stall, Corinna Gries, Nick Jarboe, Karen Stocks, Lynne Schreiber, Matt Jones

Agenda/Notes:
Next month, Meredith Goins, WDS (World Data System) on their recent review of strategic plans and technical roadmaps of WDS members

Discuss the Summer ESIP meeting:

Send out a template of talking points for the lightning talks.

The topic we decided on was for people to discuss how their facility does data quality control. This is pretty wide and can include how the facility interacts with data contributors, the data upload process, the amount and types of metadata required, metadata standards and vocabulary lists used, data validation techniques, internal data review processes, interaction with outside reviewers.

Meeting schedule (90 minutes total)
5 min intro
30 min for talks
45 min for breakouts (number determined by attendance 5-10 per group)
Group by type along the generalist to specialist data facilities
10 reconvening - groups to give a brief report on the number one point discussed (1-2min)
Mention un-conference possibilities for further discussion.

Matt Jones, DataONE, UCSB, jones@nceas.ucsb.edu, at meeting Arctic, at meeting

Corinna Gries, University of Wisconsin, Environmental Data Initiative (EDI), cgries@wisc.edu, at meeting EML

Robert Downs, SEDAC, rdowns@ciesin.columbia.edu, at meeting

Chris Jenkins, INSTAAR, Univ. Colorado Boulder, jenkinsc0@gmail.com, likely remote

Chris Crosby, OpenTopography, chris.crosby@earthscope.org , likely meeting

Danie Kinkade, BCO-DMO, dkinkade@whoi.edu, at a meeting

Session title:
Data Facilities and Data Quality: Methods of Data Acquisition, Metadata Requirements, Controlled Vocabularies, Validation and Data Review

Session Description:
This session will consist of 5 minute talks by data facility directors and staff on how their data facilities work towards archiving high data quality. Points discussed will include how data facilities interact with data contributors, the data upload process, the amount and types of metadata required, metadata standards and vocabulary lists used, data validation techniques, internal data review processes, and interaction with outside reviewers. Breakout sessions along facility types (from the more general data facilities to domain focused ones) discussing a few specific topics in detail to facilitate cross fertilization of methods and ideas. The breakout groups will then rejoin and give a brief report on the number one point discussed.

The full ESIP Session Application (submitted May 8th):
https://docs.google.com/document/d/1IBONc6V6urq3hJnyqOaeFUMALuYM9L59FgfhLdbg5Oo

NSF public access plan 2.0 recently released: https://new.nsf.gov/public-access

April 19

Attendees: Nick Jarboe, Lynne Schreiber, Corinna Gries, Matt Jones, Shelley Stall, Martin Seul, Frank Nitsche, Rob Casey

Agenda

Beyond EarthCube workshop submissions due this Friday. In Marina Del Rey (Los Angeles) June 27th and 28th.

ESIP session submissions due May 8th. What sort of session should the CDF plan for and submit? 3 min Lightning talks from various lightning talks, Nick, Matt, Corinna (data quality, meta data checks, quality analysis, work flows, FAIRness, shared vision for high quality data curation, what community needs, improving data submission)

Good to

Shelly Stall: Marian Martone - Bio and Health - Neuro repository
GREI - Gerneralist repositories.
Guidance from OSTP for repositories

Speaker topic?

Highest ranked on list:
Best practices/success stories for encouraging data citations and tracking for usage reporting

These two could be combined?

Presentations about federally supported cyberinfrastructure potentially available to data facilities: cloud storage, HPC, etc resources.

Dealing with very large datasets - growing by terabytes, accumulating petabytes. Strategies for storage and access, whether to move large data as fast as possible; keeping code close to data for compute. Possible presenters: NOAA NESDIS approach to allow scientists to run workflows on their data…what guardrails they have, etc. Globus for ideas on transport.

Notes:
Shelley proposes Marian Martone - Bio and Health - Neuro repository to speak at next CDF meeting in May

How various data repositories do data curation, methods, how to get around the difficulty of Lightning talks about what type of curation

Range of categories.
3 minutes
Matt arctic data center - Quality
Corinna Greis - will give a talk
Hydroshare - Martin
BCO-DMO
OpenDap and netCDF

Tied into the FAIR criteria.
Nelson Memo for OSTP

How to build a shared vision of what is highly useful for shared curation
What can we do together to make things better in terms of shared curation tools
To
Google

March 15

Attendees: Nick Jarboe, Karen Stocks, Frank Nitsche, Megan Carter, Lynne Schreiber, Rob Casey, Bob Downs, Chris Crosby, Amber Budden

Agenda
6 minute talk from Raleigh Martin, Program Director, Geoinformatics (GI), Division of Earth Sciences (EAR), Directorate for Geosciences (GEO), NSF from the MagIC workshop (Feb 28th, 2023). Link to talk

Topics:
Open Sciences - OSTP declared 2023 year of open science (open.science.gov)
Immediate open access plan for federally funded publications and underlying data.

NSF Public Access Repository (NSF-PAR)
PAR 1.0 peer-reviewed papers (1-year embargo period)
PAR 2.0 recently launched. PIs may now (optionally) index datasets resulting from NSF grants

Open Science opportunities across NSF
FAIROS RCNs (NSF 22-553) - Coordination of FAIR principles and open science (OS) practices. Due April 12th, 2022 (past already)

GEO Cyberinfrastructure opportunities
Geosciences Open Science Ecosystem (GEO OSE) (NSF 23-018) - Due Mar 16, 2023
AI/ML Dear Colleague Letter (NSF 23-046)
The Geoinformatics (GI) program (NSF 21-583) - Due Aug 15, 2023

Office of Advanced Cyberinfrastructure (OAC)

Cyberinfrastructure of Sustained Scientific Innovation (CSSI) (NSF 22-632) - Due Dec 1, 2023
CyberTraining (NSF 23-520) - Due Jan 18, 2024
Strengthening the Cyberinfrastructure Professionals Ecosystem (SCIPE) (NSF 23-521) - Due Jan 28, 2024

ESIP Summer (May 18 deadline), SciDataCon, ESIP session proposal, and AGU session brainstorming

Future call topic brainstorming

February 15

Attendees: Daniel S. Katz, Nick Jarboe, Christine Kirkpatrick, Karen Stocks, Ethan Davis, Frank Nitsche, Jim Riley, Leslie Hsu, Matt Jones, Simon Goring, Tamar Norkin, Yuhan (Douglas) Rao, Megan Carter, Lynne Schreiber, Kerstin Lehnert, Jerry Carter, Benjamim Branch
Agenda

CDF Governance and membership model poll (10 min)
- please vote now if you have not already
- results:
    - The membership decided to keep the past institutional membership model, (65%)
    - The membership model does not impact my attendance (78%)
    - Meeting announcements and presentations will be recorded, but not discussions (50%), (whole meetings 44%)

Presentation on The FAIR in ML, AI Readiness, & Reproducibility Research Coordination Network (FARR RCN; www.farr-rcn.org). (10 min)

FARR aims to build better practices (via a roadmap, community practices, and advice on tooling) for both the members of and the larger CS, GEO, and other communities they represent. This will lead to products (e.g., data, models) that are more FAIR, which in turn will lead to greater reproducibility where these products are used, and increased reuse of the products.
Slido poll results:
Responses to AI/ML goals:machine-readable data, harmonized large-volume data stores (synthesis of long-tail data

Hired data scientist to develop initial AI/ML tools
Provide direct access to data in our repository (s3 buckets) through a common format
In process of moving our data and applications to the cloud, our systems should then start to be more AI/ML ready. But that would be a stretch goal.
Some include: 1) improve semi-automated workflows for creating AI-ready input data for training, testing, and model application; 2) improve workflows for AI-output post-processing, especially preparation for visualization of large data outputs; 3)

understand unique metadata and documentation needs for AI data and models for preservation and sharing
Mainly around training for our community.
Get a shared understanding of what that means.

Discussion of AI-readiness interests, needs, and progress in the CDF community (25 min)

Future Meeting Topics (10 min)

Notes:
May need more than FAIR data
When the community starts using ML/AI then we know we are ready
There are different ML/AI applications and may be different for readiness
Need for best practices for tooling
There is Problem with a pipeline problem-human training
Those already doing AI mostly need just access
Does the prepared data or models ever come back to facilities for others to use? - at IEDA they do the processing of data for the researchers, place for code to be re-used
Is there a need for others (not using Ai/ML) to see what is possible - more skilled seem needed, facilities are running virtual workshops

Next meeting on March 15

January 18, 2023

Attendees: Nick Jarboe, Lynne Schreiber, Karen Stocks, Corinna Gries, Bob Downs, Doug Fils, Lindsay Powers, Frank Nitsche, Kristina Vrouwenvelder, Michael Bobak, Danielle Lopez, Matt Jones, Martin Seul, Megan Carter, Danie Kinkade

Agenda:
Structure of CDF cluster going forward
Expectations of cluster
Minimum requirements to be an active member
CDF Chair rotation schedule

General assembly

Topics for upcoming meetings

Notes:
Past Survey of the CDF deciding to move to a ESIP Cluster

- Should CDF maintain a formal membership?
    - M. Jones: What would the benefits be or how would this barrier help us?
    - N. Jarboe: easier to attribute credit/agreement/ group consensus

    - K. Stocks: others would still be welcome to join and participate of course
    - Lots of questions about how this relates to ESIP partnership, whether being an ESIP partner is enough.
    - M. Seul raised that there is value in having a formal membership still, that for him it does make a difference
    - Others agreed that there is a difference between showing up as an individual or showing up as a designated representative from a repository.
    - Chris Crosby - year long process to get funding from NSF and USGS.
    - Send out link to the CDF members with link to the topic list and suggest comments or presentations.
    - .

Pros/cons/issues for more formal membership:

- Pro: Sense of community/belonging
- Pro: Cohesive voice for messaging
- Pro: commitment to group
- Con?: more difficult to become involved, creates a barrier for participation
- Con?: lots of effort to set up and maintain memberships if we don't really need one
- Issue: Can individuals be members?

Have a poll on which model is preferred, and perhaps whether one model would make it more likely for members to attend. Draft wording:

- Institutional membership with designated representatives for each institution; formal application & review; only members vote. But mailing lists, meetings and other activities open to all.
- Open individual participation with no application; all members can vote

Note that votes are used for changes to cluster terms of reference, any elected position (there currently aren't any), membership decisions. In theory, votes can be used to endorse consensus statements, white papers, etc. but historically this has not yet been done.

Also note that terms can be changed in the future.

And an up-down vote on a 2-year rotating chair an co-chair

Maybe add a poll about which membership model would most facilitating that person's future involvement

And a question about recordings - 3 options: no recording, recording only announcements and presentations, or record everything. Just for member use.