CDI Monthly Meeting 20211110

November 10, 2021: Data standards primer, Text tools for automation and reproducibility, Statistics challenges, Confluence wiki migration

The Community for Data Integration (CDI) meetings are held the 2nd Wednesday of each month from 11:00 a.m. to 12:30 p.m. Eastern Time.

Meeting Recording and Slides

2021-11-...ides.pdf

These are the publicly available meeting materials. Log in to view all meeting resources. If you would like to become a member of CDI, join at https://listserv.usgs.gov/mailman/listinfo/cdi-all.

During the call, you can ask and up-vote questions at slido.com, event code #CDINOV.

Agenda (in Eastern time)

Times are approximate

11:00 am Welcome and Opening Announcements

11:15 am CDI Collaboration Area Updates

11:25 am **Disciplinary data standards: an example and primer** - Abby Benson, USGS

11:35 am **Text tools for automated and reproducible research: Markdown and LaTeX** - Richie Erickson, USGS

11:45 am **Statistics challenges**

11:55 am **Highlights from recent data-related meetings**

12:05 pm **Confluence wiki migration**

12:30 pm *Adjourn*

Abstracts

Disciplinary data standards: an example and primer

Abby Benson, USGS

The ESIP Biological Data Standards Cluster has been developing a primer on existing biological data standards for managers of biological data who may be unaware of existing standards but need to improve management, analysis, and use of the biological observation data. The goal of this primer is to spread awareness about existing standards in a simple, aesthetically pleasing way. Our hope is that this primer, shared online and at conferences, will help increase the adoption of existing biological standards and help make data more Findable, Accessible, Interoperable, and Reusable (FAIR). (From Biological Data Standards - A Primer for data managers)

Benson, Abigail; LaScala-Gruenewald, Diana; McGuinn, Robert; Satterthwaite, Erin; Beaulieu, Stace; Biddle, Mathew; et al. (2021): **Biological Observation Data Standardization - A Primer for Data Managers**. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.16806712.v1

Abby Benson is a biologist in USGS Science Analytics and Synthesis (SAS).

Text tools for automated and reproducible research: Markdown and LaTeX

Richie Erickson, USGS

Markdown (https://daringfireball.net/projects/markdown/) and LaTeX (https://www.latex-project.org/) are both mark-up languages allowing users write in plain text and how the language compile and format documents. These tools allow USGS scientists to be more efficient by automating portions of their research and more reproducible by linking models directly to outputs. Historically (ca, 1980) and to the present, scientists and mathematicians have used LaTeX when typesetting equations because word processors lacked the ability to correctly typeset equations. More recently (ca 2000), scientists and mathematicians have used LaTeX to embedded code and code outputs directly into self-updating documents. Markdown exists as an easier-to-use language than LaTeX and, currently (ca 2012), Markdown may be used to generate documents with code and results embedded directly within them.

Richie Erickson is a Research Quantitative Ecologist at the Upper Midwest Environmental Sciences Center.

Highlights

1. Vote on statements of interest by midnight on November 12th!
2. Please fill out this survey about your use of USGS Metadata Tools: https://forms.office.com/g/fq64L4htj9
3. Check out the new Biological Data Standards Primer: https://doi.org/10.6084/m9.figshare.16806712.v1
4. Interested in Markdown and LaTeX? Join the new group here: https://forms.office.com/g/M6522ybAAj
5. Statistics course are available!
    a. Statistical Techniques for Trend and Load Estimation from Jan 24-28, 2022.
    b. Statistical Methods for Environmental Data Analysis from Mar 7-11, 2022.
    c. Contact: kryberg@usgs.gov
6. Discuss SharePoint solutions together: https://doimspp.sharepoint.com/:x:/s/CommunityforDataIntegration/EQHJ1Zw_hrxOprri-tWSeqIBAZruE_JOnbpIK3zKfAzHLQ?e=rFqXGp

Welcome and Opening Announcements

1. New census site https://www.census.gov/about/what/data-equity.html
    a. Related to our 'Advancing Data Equity' FY22 proposal theme
    b. https://www.census.gov/about/what/data-equity.html (short 10 minute video on community resilience)
2. Past CDI project sightings
    a. mdEditor – still around and available for ISO metadata
    b. Data Management Training Clearinghouse – mentioned in the Research Data Alliance (RDA) meeting
    c. FAIR data roadmap for USGS – mentioned in the RDA meeting
3. Vote on statements of interest by midnight on November 12th!
    a. https://doimspp.sharepoint.com/sites/CommunityforDataIntegration/Proposals/Proposal_Submissions/Forms/FileView.aspx?viewid=8c6f8412%2Dc261%2D4f8e%2Da03f%2D98d9008935ce&OR=Teams%2DHL&CT=1634590407837http%3A%2F%2F

CDI Collaboration Area Updates

1. For more information on any of the collaboration areas, see https://my.usgs.gov/confluence/x/yhv1I
2. Risk
    a. **Risk Research and Applications RFP now open:**
    b. https://my.usgs.gov/confluence/display/cdi/Risk+Research+and+Applications+FY22+RFP+Application+Materials%2C+Resources%2C+and+FAQs
3. Usability
    a. Next event: December 1st : **Selecting discussion/presentation topics for 2022**
    b. November resource review: User Interface Design Elements
4. Data Management
    a. Next event: December 13th: GeoPackage Overview and USGS Use Cases
    b. *****Please fill out survey** about your use of USGS Metadata Tools***: https://forms.office.com/g/fq64L4htj9
5. Inland/Coastal Bathymetry
    a. Next event: **November 23rd**
6. Geomorphology
    a. Next event: **November 16th**: USGS Bathymetric and Topobathymetric Data Inventory
7. Data Viz
    a. Next event**: December 2nd**: NASA-SERVIR, Africa Flores
8. eDNA
    a. Next event: TBD
    b. Past event: The Government eDNA working group's 5th annual eDNA Technical Exchange Workshop
        i. **Recorded talks available through the wiki**
    c. Imagery
        i. Next event: Imagery Data Management Workflow demo, TBD
        ii. **Camera-based monitoring workshop, two half-day sessions, planned for Jan/Feb 2022**

Disciplinary data standards: an example and primer - Abby Benson, USGS

1. Many repositories and groups were wondering what standards people are using for biological data, group produced a primer, "Biological Observation Data Standardization: A primer for data managers"
2. The primer covers:
    a. **Why you should use standards (even if you're not sharing your data)**

   b. Questions a data manager may ask (i.e., Do you want to… provide context and understandability to your data?), with answers and applicable standards
   c. Acronyms are clickable links!
   d. The participants in creating this primer were mainly in the marine sector, would like more freshwater or other scientists
3. What's next?
   a. The cluster is using the primer and moving forward with it. We're working through use cases (bringing a biological observation dataset to the cluster, using this primer to see if questions come up).
   b. **The cluster would love more biological observation datasets for use cases.**
4. See the primer here: https://doi.org/10.6084/m9.figshare.16806712.v1

Text tools for automated and reproducible research: Markdown and LaTeX - Richie Erickson, USGS

1. Problems
   a. How do I create reproducible results?
   b. How do I embed data/code output files?
   c. How do I typeset complex math?
   d. Quickly change formatting such as journal styles?
   e. Write a large document such as a dissertation
2. Solutions
   a. LaTEX and markdown
      i. Markdown uses LaTeX to create PDFs
      ii. Markdown also does more: HTML, eBooks, Word

- RMarkdown replaces SWeave

1. Who uses LaTeX or Markdown in USGS?
   a. All GitHub/GitLab users for README.md files
   b. Earthquake hazard reports (get outputs from instruments, automatically generate a report)

- Water Mission Area reports and documentation

1. LaTeX for journal articles
2. RMarkdown for journal articles, great way to embed code
3. Brave few use LaTeX for SPN products

1. How to use in USGS?
   a. RStudio for RMarkdown/LaTeX
   b. Pandocs (program) for general Markdown

- TeXLive for LaTeX

1. Jupyter Notebooks – use markdown for text blocks

1. Markdown/LaTeX group
   a. See who is using tools in USGS
   b. Identify group needs

- Advocate for our needs

1. Co-chairs: Richie Erickson and Leslie Hsu
2. Interested? Join here: https://forms.office.com/g/M6522ybAAj

1. CHS is hosting a RStudio presentation and demo:
   a. RStudio Team User Group Meeting - Next Thursday, November 18th at 2 PM ET. Our representative from RStudio will give an introduction and demonstration of RStudio Team.
   b. This will be hosted in the GS-CHS-UserCommunity Microsoft Team; it is recommended to join the team prior to the meeting to avoid connection issues.

Statistics challenges

1. Statistics course are available:
   a. **Statistical Techniques for Trend and Load Estimation from Jan 24-28, 2022.**
   b. **Statistical Methods for Environmental Data Analysis from Mar 7-11, 2022.**
   c. Contact: kryberg@usgs.gov

Highlights from recent data-related meetings

1. Fish & Wildlife Service Data Management Workshop
   a. Tidy data, plenary sessions, and implementing quality assurance on the website:
      i. October 2021 FWS Data Management Workshop: https://doimspp.sharepoint.com/sites/fws-data/SitePages/October2021DMWorkshop.aspx
   b. FWS Data Management SharePoint site includes data management resources: https://doimspp.sharepoint.com/sites/fws-data
2. Virtual SciDataCon 2021
   a. Video of Making Your Data Center and Services Ready for AI: Case Studies at https://vimeo.com/637313425.

3. Research Data Alliance
    a. 3-11 November 2021
    b. Many interest groups and working groups
        i. RDA Sensitive data Interest Group: https://www.rd-alliance.org/groups/sensitive-data-interest-group

Confluence wiki migration

1. **Our access to Confluence wiki is ending in early 2022**
2. CDI has ten years of content on the wiki
3. We are in the process of migrating our content
4. Purpose of the space
    a. Storage (slides, notes, recordings)
    b. See what’s coming up (calendars, agendas)
    c. Work together on things like the ongoing Proposals process
    d. See recording for demo of space, https://doimspp.sharepoint.com/sites/CommunityforDataIntegration
5. **Discuss SharePoint solutions together**: https://doimspp.sharepoint.com/:x:/s/CommunityforDataIntegration/EQHJ1Zw_hrxOprri-tWSeqIBAZruE_JOnbpIK3zKfAzHLQ?e=rFqXGp

Questions

1. Data Standards Primer
    a. What does "Structure data in long format" mean?
        i. Some people put things that should be in rows as a column header (like species name); we want it in long format rather than wide.
    b. With multiple standards available for data, metadata, etc., do we have to choose one or support all? If choose one, what's the criteria for making the choice?
        i. Some people in the cluster favor one standard over another. But how do you make that choice? You may have a place you’re aiming to share the data and would want to adhere to their standards.
    c. Is there inconsistency in using standards internationally?
        i. There is not a lot of standardization of biological data, internationally. This is part of the reason this cluster was stood up.
    d. Markdown/LaTeX
        i. Using Markdown and LaTeX to generate PDFs: can it generate a structured, tagged 508-compliant PDF?
            1. I know this has been a conversation, but don’t know the answer.
        iii. What’s the best way to track changes in Markdown/LaTeX?
            1. It depends on who you’re tracking changes for. For BAOs/journal articles, you can compile TEX documents and shows the differences. Otherwise, I use git tracking.