Outcomes from the Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons

Doug Schuster, *Matt Mayernik, John Cline*
NSF National Center for Atmospheric Research
Deanesh Ramsewak
University of Trinidad and Tobago

Jan, 2025

This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

Presentation Outline

- Workshop motivation -Challenges with existing data services
- Emerging data services
  - In the community
  - At NSF NCAR
  - Global scale
- Workshop recommendations

About NSF National Center for Atmospheric Research

Our eight labs and programs collectively cover a breadth of research topics in Earth system science, from the effects of the Sun on Earth's atmosphere to the role of the ocean in weather and climate prediction, as well as supporting and training the next-generation of Earth system scientists.



EDUCATION, ENGAGEMENT & EARLY-CAREER DEVELOPMENT (EDEC)

EdEC is committed to inspiring, engaging and informing the public about the science going on at NCAR and in our broader scientific community.

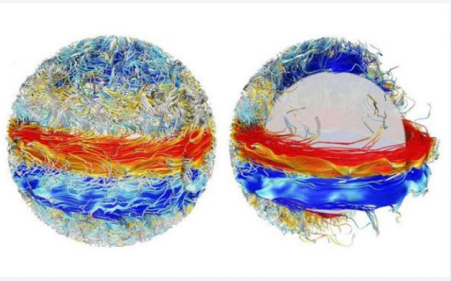
VISIT EDEC



EARTH OBSERVING LABORATORY (EOL)

EOL provides state-of-the-art atmospheric observing systems and support services to the university-based research community for climate, weather, and related Earth system research.

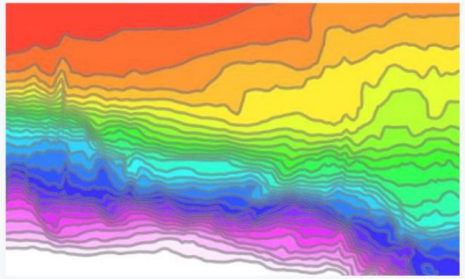
VISIT EOL



HIGH ALTITUDE OBSERVATORY (HAO)

HAO conducts fundamental and applied research in solar-terrestrial physics using observational, theoretical, and numerical methods. Research at HAO extends from the solar core to the surface of the Earth.

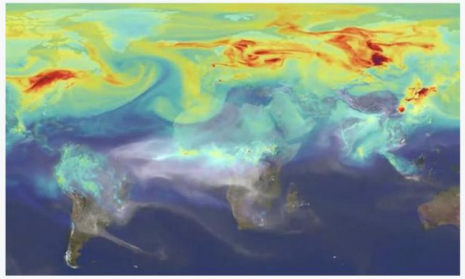
VISIT HAO



ATMOSPHERIC CHEMISTRY OBSERVATIONS & MODELING (ACOM)

ACOM research focuses on advancing understanding and predictive capability for atmospheric composition and related processes.

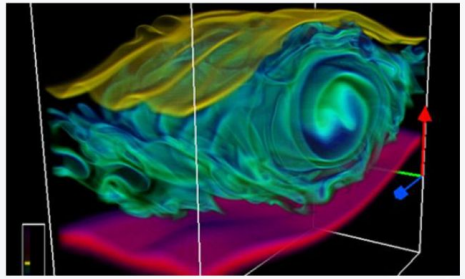
VISIT ACOM



CLIMATE & GLOBAL DYNAMICS (CGD)

CGD develops understanding of the Earth system and is a leader of community-developed and community-owned models of the Earth system.

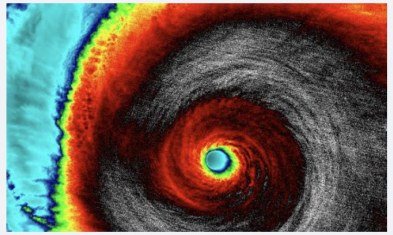
VISIT CGD



COMPUTATIONAL & INFORMATION SYSTEMS LABORATORY (CISL)

CISL is a world leader in supercomputing and cyberinfrastructure, providing services to NCAR, UCAR member universities, and the broader geosciences community.

VISIT CISL



MESOSCALE & MICROSCALE METEOROLOGY (MMM)

MMM advances the understanding of regional and fine-scale aspects of weather and climate and applies this knowledge to benefit society.

VISIT MMM



RESEARCH APPLICATIONS LABORATORY (RAL)

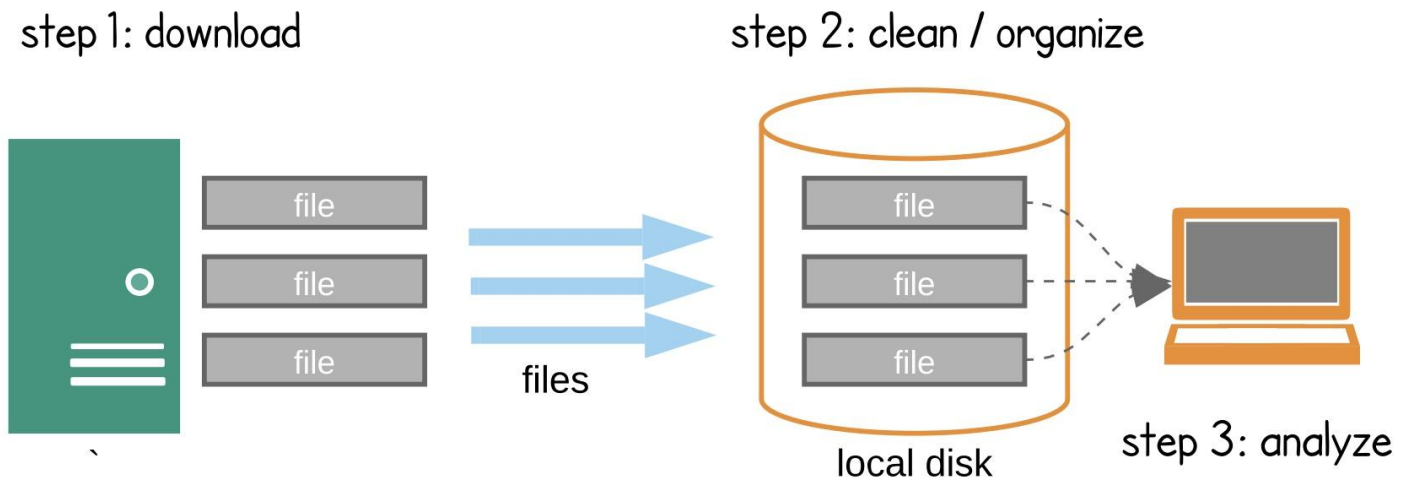
RAL conducts directed research and engineering to discover solutions to problems relevant to society and facilitates the transfer of our information, expertise, and technology to the public and private sectors.

VISIT RAL

Current Data Services Challenges

Remote Users -> Legacy download and analyze model

- Time consuming/inefficient
- Not feasible to for most users to download and store 100s/1000s TB
- Inconsistent use of formats, conventions and structures creates challenges
- Differing user experiences across NSF NCAR's repositories



Credit: Ryan Abernathy

Current Data Services Challenges

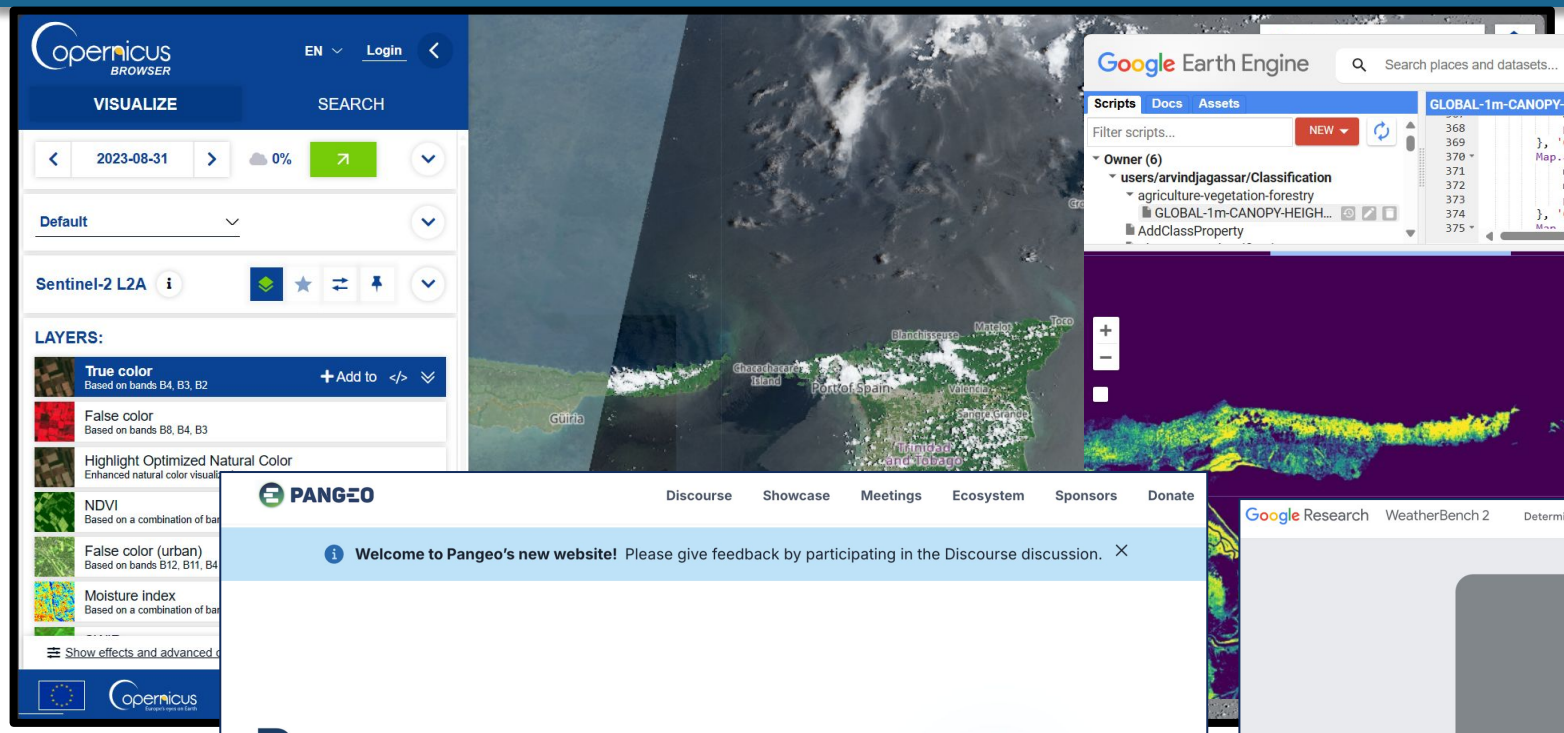
NSF NCAR compute services users -> Direct data read access

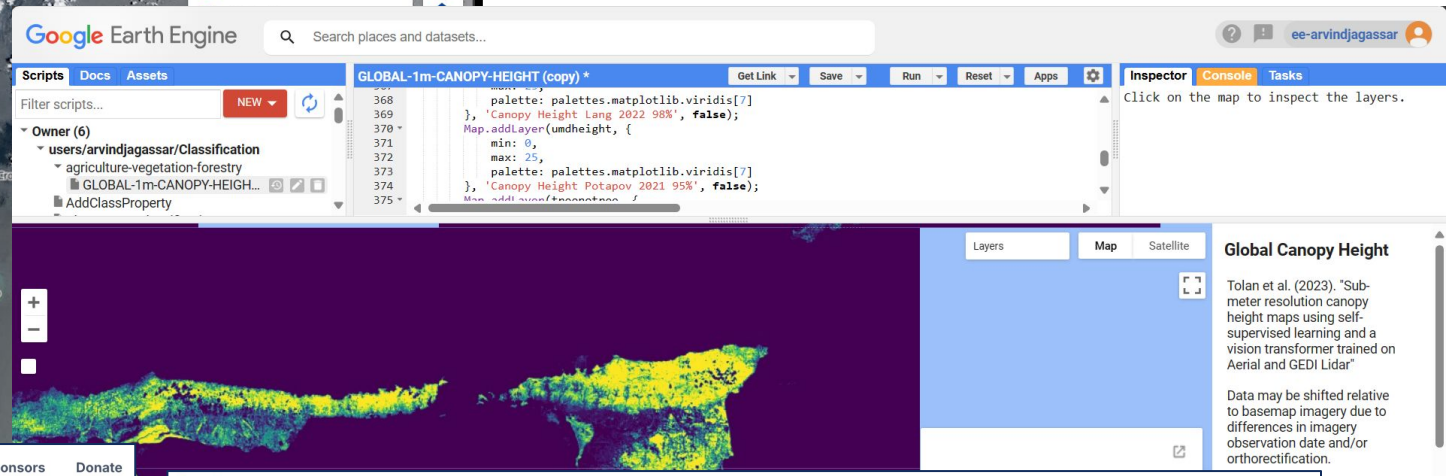
- Limited to users with NSF NCAR compute services accounts/allocations
- Insufficient search capabilities, data structures complex to access
- Not of all of NCAR's data are accessible from these systems

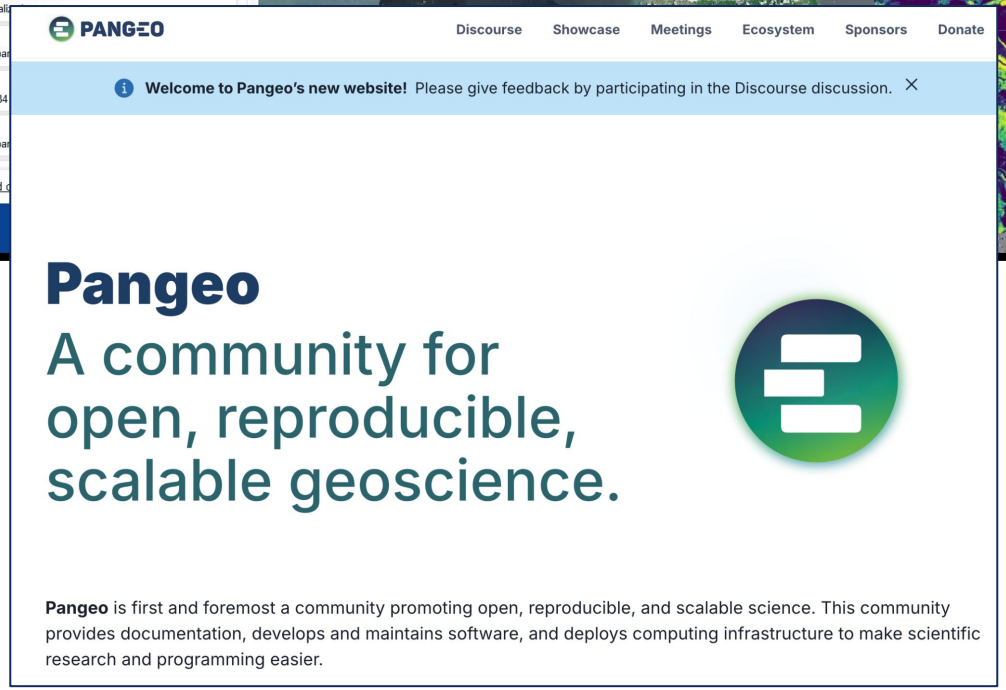


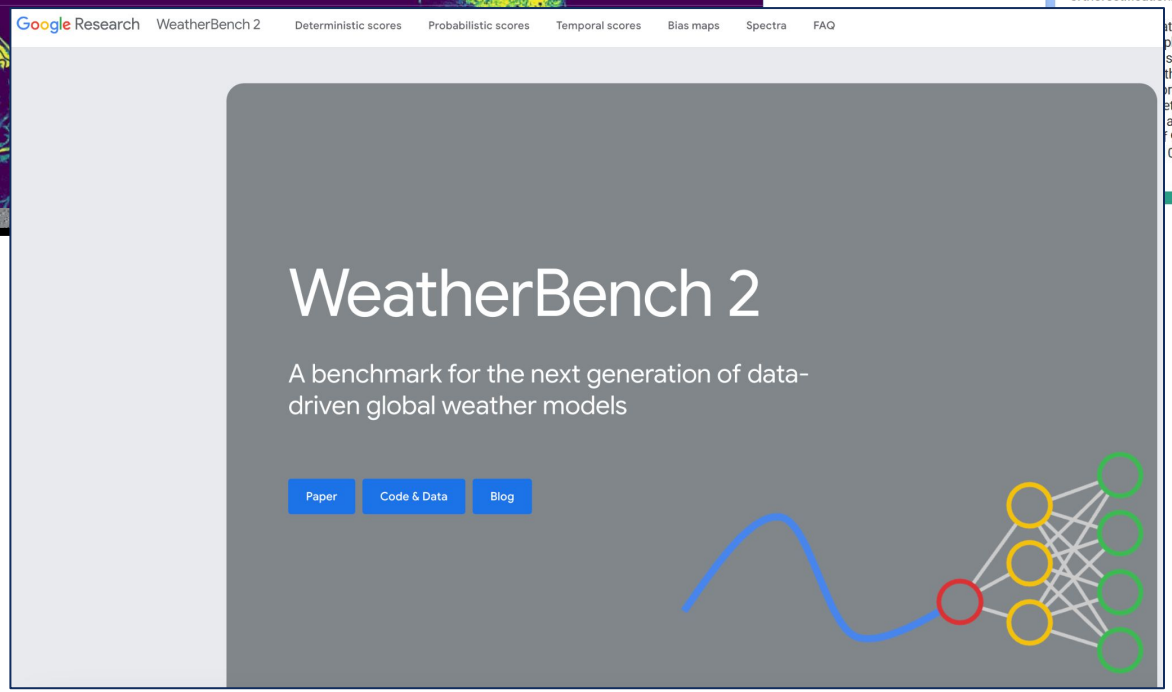
- Retrospective Atmospheric Reanalyses
- Observations (Global and Field Campaign)
- Climate Model Simulation Future Scenarios

Emerging Data Services -Traditional, Private Sector and Community Solutions Built on Commercial Clouds









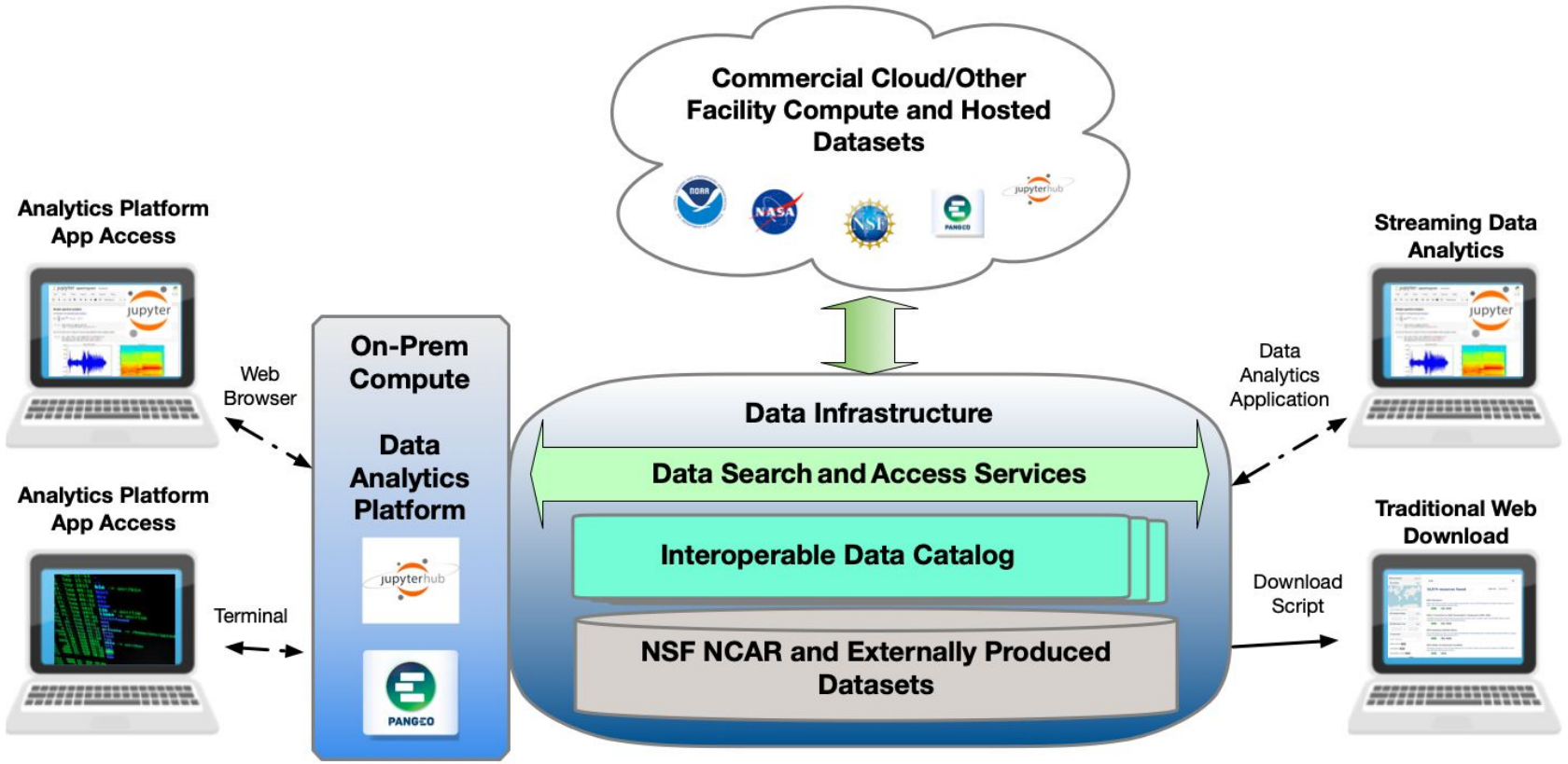
Pivoting to NSF NCAR's Integrated Research Data Commons

What is our vision for an Integrated Research Data Commons?

Data science infrastructure that connects geoscience data sets with community developed analytics tools to allow users to share, integrate, analyze, and visualize geoscience research data to drive scientific discovery.

Grossman, R.L. Ten lessons for data sharing with a data commons. *Sci Data* 10, 120 (2023). https://doi.org/10.1038/s41597-023-02029-x

Pivoting to NSF NCAR's Integrated Research Data Commons



Realizing the potential of NSF NCAR's research data

- **Provides entry points for novice to experienced researchers**
- **Incorporates integrated tools for data exploration, analysis, data assimilation, and AL/ML research.**
- **Positions NSF NCAR to support emerging International Scale Cyberinfrastructure initiatives**

Grossman, R.L. Ten lessons for data sharing with a data commons. *Sci Data* 10, 120 (2023). https://doi.org/10.1038/s41597-023-02029-x

Emerging Need: Integrate with Global Scale Cyberinfrastructure Initiatives

The Open Science Data Federation is a federated platform for delivering datasets from repositories to compute in an effective, scalable manner.

- **Streaming data delivery/access -Compute location agnostic storage**



Community Expectations for a Geoscience Data Commons Workshop

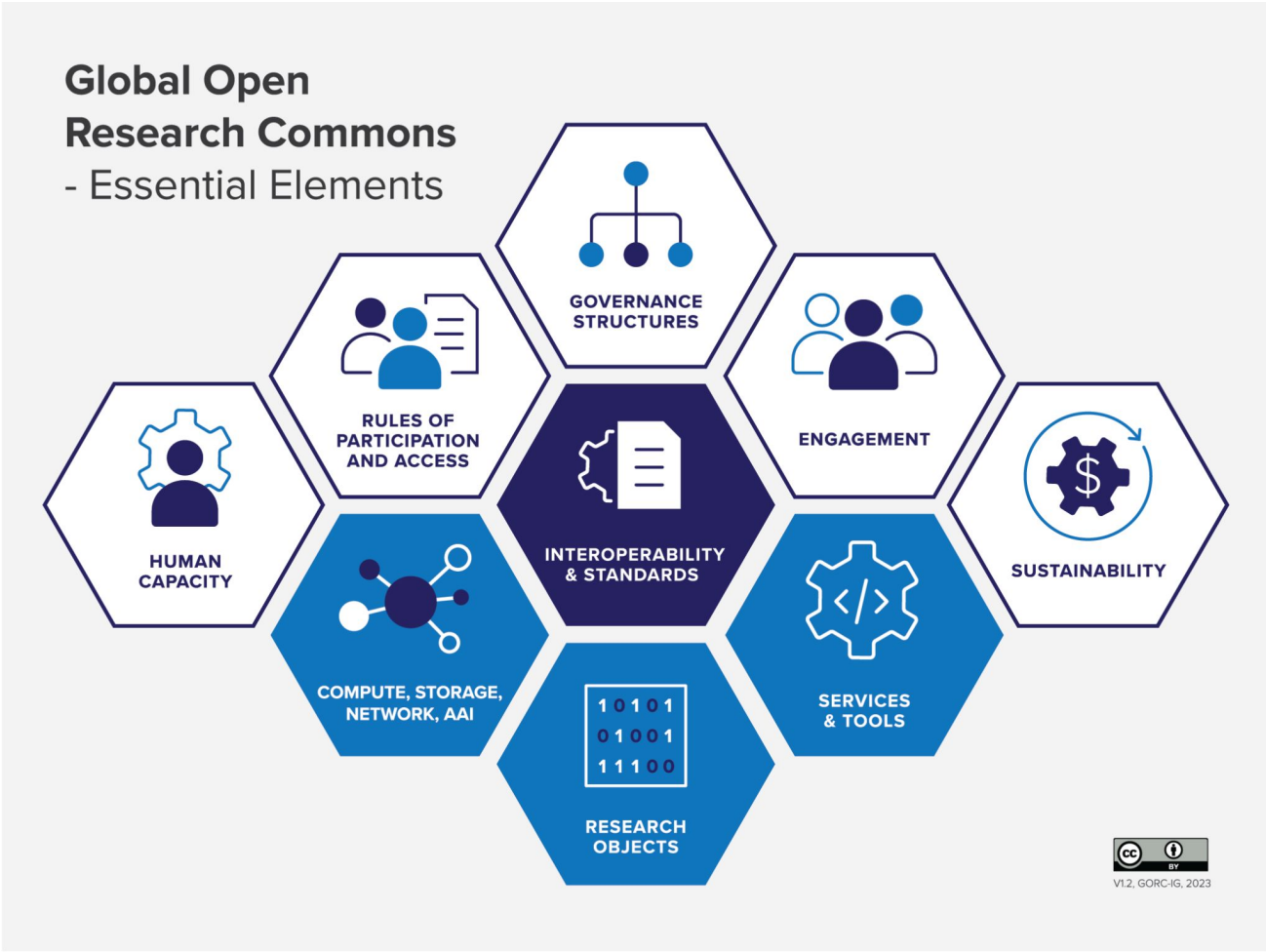
- May 29-31, 2024 in Boulder, CO
- 70 participants
  - International
  - Universities
  - Cross Agency
  - 15% Representing Minority Serving Institutions



Community Expectations for a Geoscience Data Commons Workshop

- **Workshop Report**:
  Mayernik, M. S., Schuster, D. C., & Clyne, J. P. (2024). Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons - Workshop Report (No. NCAR/TN-584+PROC). doi:10.5065/gfbq-8y08

- **Recommendations aligned with the concepts of the Global Open Research Commons**



Woodford, C., Treloar, A., Leggott, M., Payne, K., Jones, S., Lopez Albacete, J., Madalli, D., Genova, F., Dharmawardena, K., Chibhira, N., Åkerström, W. N., Macneil, R., Nurnberger, A., Pfeiffenberger, H., Tanifuji, M., Zhang, Q., Jones, N., Sesink, L., Wood-Charlson, E., & RDA GORC International Model WG. (2023). The Global Open Research Commons International Model, Version 1 (1.0). Zenodo. https://doi.org/10.15497/RDA00099

Selected Workshop Findings

- **Governance Structures**
  - A community informed governance strategy is needed to develop policies to structure processes for data deposit, documentation, access, retention, ***deletion***, and preservation.
- **Engagement**
  - Data commons developers need to be aware of ***multiple entry points for novice, intermediate, and advanced users***, and develop comprehensive ***community engagement strategies to ensure that data commons services respond to the needs of the intended user communities***.
- **Research Objects**
  - A mechanism should be provided that allows ***community feedback for datasets and their metadata*** (e.g., similar to GitHub issues).
- **Services and Tools**
  - A data commons resource should ***facilitate creativity and innovation by its user communities*** vs limiting them to a predefined set of applications
    - For example, allow researchers to bring their own container to a commons environment and share code with one another in a searchable way.

Selected Workshop Findings

- **A data commons should be designed to satisfy tractable goals while also enabling less-advanced communities.** Participants generally agreed that a data commons should be a resource that ***unlocks the untapped research potential of currently siloed***, ***domain-science research datasets***, while also broadening the community of users who can participate in data science research.
- **What is the common denominator for a data commons?** It was agreed upon that a data commons ***needs to be built for and by its user community/people, and be a gathering place for its user community to co-mingle and exchange research artifacts*** (e.g., data, code, and ideas).
- **Trust as a community is paramount for success.** There was much discussion amongst participants that ***different communities need to own their own products*** (e.g., data sovereignty -similar to the concepts described in the CARE principles for Indigenous Data Governance) to instill trust in their downstream user communities and stakeholders.

Questions?
schuster@ucar.edu

