Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons - Workshop Report

Matthew S. Mayernik
Douglas C. Schuster
John P. Clyne

NCAR Technical Notes
NCAR/TN-584+PROC

National Center for
Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR TECHNICAL NOTES

http://library.ucar.edu/research/publish-technote

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not yet at a point of a formal journal, monograph or book publication. Reports in this series are issued by the NCAR scientific divisions, serviced by OpenSky and operated through the NCAR Library. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution maybe limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307-3000

NCAR/TN-584+PROC

NCAR Technical Note

2024-08

Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons - Workshop Report

Matthew S. Mayernik
NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA

Douglas C. Schuster
NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA

John P. Clyne
NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA

NCAR Library

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

How to Cite this Document:

Mayernik, Matthew, Douglas Schuster, John Clyne. (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

Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons -Workshop Report

Matthew S Mayernik
Doug Schuster
John Clyne

Acknowledgements

This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977, in response to the US National Science Foundation Dear Colleague Letter (NSF 23-141): “Innovations in Open Science (IOS) Planning Workshops”. We thank Taysia Peterson and the NSF NCAR administrative and technical support staff for their support in organizing the workshop, and all of the workshop attendees for their participation and feedback on earlier versions of this report. We also thank the members of our workshop steering committee (listed in Appendix I), and in particular Rebecca Ringuette for serving as a reviewer of this Tech Note.

Table of Contents

1. Executive Summary

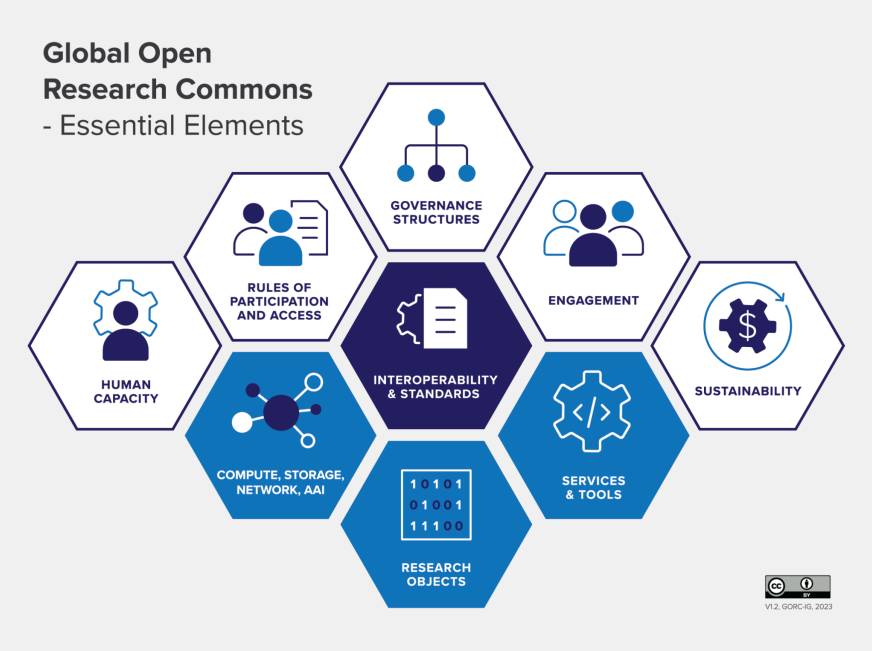
The National Science Foundation Division of Atmospheric and Geospace Sciences (NSF-AGS) and the NSF National Center for Atmospheric Research (NSF NCAR) funded Innovations in Open Science (IOS) Planning Workshop: Community Expectations for a Geoscience Data Commons was held from May 29-31, 2024 in Boulder Colorado. The workshop brought together a diverse group of over 70 participants to provide a variety of perspectives on community needs for a geoscience research data commons environment. Stakeholder communities represented at the workshop included geoscience researchers and students (weighted towards the Atmospheric, Hydrologic, Geospace and Oceanic sciences) from US Universities (including Minority Serving Institutions (MSIs) and Historically Black Colleges and Universities (HBCUs)), technology experts, representatives from open source communities of practice such as Pangeo [1], private sector organizations, and data repository personnel. The overall goal of the workshop was to ask participants to develop recommendations to satisfy community needs for geoscience research data commons from a broad range of perspectives. Additionally, we wanted to build a welcoming and inclusive community through the workshop that will provide the foundation for continued collaborative activities that result from the workshop.

The workshop agenda was organized with three predefined plenary and breakout discussion session themes, that included: 1) Big ideas and existing solutions for data services and data analytics capabilities, 2) Data commons analytics research and educational user needs, and 3) data producer needs for data curation and consulting services. For each of these themes there was a plenary session that included 5 topical talks to promote ideas and stimulate discussion in breakout sessions that followed. The breakout sessions included 7 separate breakout groups that workshop participants were randomly assigned to, and each of these groups discussed identical questions for these first 3 workshop themes. A 4th breakout session took place during the last day of the workshop. For this session, workshop participants were given the opportunity to suggest and vote on breakout group themes that they felt would be of value for further discussion.

A number of recommendations were developed as a result of the workshop discussions. The recommendations are organized around the components of the Global Open Research Commons (GORC) model shown in Figure 1, which was introduced at the start of the workshop and provides a robust framework for structuring the topics that were discussed during the workshop.

[1] https://pangeo.io/

Figure 1 - Global Open Research Commons - Typology of Essential Elements



A summary of workshop recommendations organized around the Global Open Research Commons Essential elements follows.

- **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.
  - Engagement strategies for a data commons must include outreach to all types of users, especially including traditionally underrepresented communities from non-R1 universities, MSIs, HBCUs, and other institutions that may have limited resources.
- **Rules of Participation and Access**
  - A data commons must have well-defined rules, policies and licenses describing what users are able to do with its data, software, and services. These rules and policies should define boundaries around usage, including deposit, storage, and access of data, and access and usage of software or computing services.
    - Guidance should be provided to data depositors on what data they need to preserve and share and for how long (e.g., raw observations, statistically processed observations, model outputs).

        - Tiered levels of value added services could be made available and provisioned through an equitable allocations process, similar to the NSF ACCESS[2] strategy.
            - Value added services, such as hosted computational notebook resources, could especially open up access to data science research for traditionally underrepresented communities from non-R1 universities, MSIs, HBCUs, and other institutions that may have limited resources.
        - Policies for the use and management of resources (data & infrastructure) need to be clearly stated to user communities, and technical capabilities should ideally be provided to enforce such policies.

- **Human Capacity**
    - There is a strong desire for sponsor-supported centralized data curation resources, similar to the NASA Data Repositories dataset assignment model [3]. In this model, the funder provides a designated repository where datasets that result from funded projects can be deposited along with a chosen level of curation support. Such a service will need well defined expectations and scope of service. The benefits of such a service include:
        - Data providers could focus their efforts on data collection and analysis.
        - Data management planning and implementation would be more straightforward
        - Oversight in the proposal review process would be easier
        - Support for data management & preservation could begin at the proposal stage
- **Sustainability**
    - A data commons needs a business model to support sustained infrastructure, including the people, storage, computing resources, and protocols involved.
    - There must be a long and predictable funding lifecycle for a data commons to be sustainable and meet community expectations for being a trusted service.
- **Interoperability and Standards**
    - Data and metadata standards and best practices on using them need to be determined by domain science communities in conjunction with data commons personnel, followed by work toward crosswalks and interoperability tools across those community solutions.
    - Tools and consulting support needs to be provided to abstract out the details of metadata development for data depositors.
    - A data commons should provide standards-based APIs to abstract out complexities and enable simplified data and metadata access.
    - Data Commons should strive to maximize interoperability with peer systems at the national and international level (e.g., NASA, NOAA, NSF, ESA, Copernicus, etc..).

[2] https://access-ci.org/
[3] https://pub.earthdata.nasa.gov/data_publication_guidelines

- Broad and active participation on standardization forums (e.g., Open Geospatial Consortium) at the national and international level will facilitate wider interoperability and synergies with data commons systems worldwide (international and cross-domain).

➢ **Compute, Storage, Network, AI**

- Data storage and access solutions, whether based on cloud services or other technical approaches, must ensure that both short and long term data preservation and sharing are supported.
- Storage, access, and analysis services must also scale appropriately for different volumes of data, and ideally storage will be performantly accessible from many types of compute services to avoid the need for replication of datasets.
- Leverage the capabilities provided by Large Language Models to:
  - Simplify data discovery and access
  - Automate metadata extraction
- Provide access to community developed, lightweight, pre-trained, and tunable machine learning models.
- Data commons developers must envisage technical capacity for increasing user demand and account for cybersecurity considerations.
- Data commons should be architected with a flexible cyberinfrastructure that will allow for the migration of data & services between differing platforms and infrastructures.

➢ **Research Objects**

- Minimum requirements and the purpose for dataset metadata should be well described (e.g., support search, provenance and modern data science needs).
  - Metadata generation and provenance tracking should be automated where possible.
  - The metadata requirements should vary depending on the properties and expected impact of the dataset.
  - Persistent Identifiers should be assigned to dataset landing pages and possibly individual data objects where appropriate
- 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.
- Integrate open source community developed software solutions into data analysis and access tools.
- Pursue public/private sector collaboration and partnerships to keep up with the technological pace of change and deliver modern/current services and tools.

Finally, a number of recurring topics came up during both plenary and breakout session discussions throughout the workshop, but did not necessarily fit into one of the essential elements of a research commons model.

- **Plenty of gaps exist in data service needs.** While many cylinders of excellence exist to provide data services, workshop participants noted that numerous gaps still exist in the repository ecosystem for supporting publisher data sharing requirements, and this leads to many questions including, "Where should my data be archived?", "What capabilities do data repositories provide?", "Who pays for long-term data curation?", and "How do we avoid silos?".
- **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).
- **Culture change is needed along with (perhaps new) business models.** Participants agreed that incentive structures in academia need to change to reward activities that support data curation and open science needs vs hoarding of resources by individual groups to produce as many peer reviewed publications as possible.
- **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 [4]) to instill trust in their downstream user communities and stakeholders.

[4] https://www.gida-global.org/care

2. Introduction and Background

Data intensive research, including data analytics, machine learning, and data assimilation continues to drive innovation and discovery across the geosciences. Geoscience datasets maintained in domain repositories, such as climate model projections, historical reanalysis products and observing facility produced datasets provide rich resources to support these initiatives. Presently, these datasets are primarily maintained in disconnected, domain-focused data systems designed to support the legacy “download, clean and analyze model” that requires data to be downloaded to a local system and reorganized before any analysis can happen. This is a time consuming process with bandwidth and storage requirements that may be prohibitive, particularly for less resourced institutions. Additionally, datasets maintained through these domain-focused repositories may be provided with unique naming standards, metadata, data structures, and formats that are likely to create barriers to access for those outside of a domain, including those investigating interdisciplinary research ideas and attempting to generalize computational workflows. This combination of the download, clean, and analysis model and the use of non-standard formats combine to create a barrier to exploiting the full research potential of geoscience data assets.

This report describes outcomes of a two day workshop that brought together more than 70 diverse individuals from multiple stakeholder groups to identify and outline requirements for a geoscience research data commons environment. Workshop participants participated in plenary sessions and breakout discussions that focused on developing requirements to modernize community-accessible data science infrastructure, to better connect our geoscience datasets with geoscience-focused analytics environments, and to support researcher needs in meeting data sharing expectations in alignment with the FAIR (Wilkinson et. al., 2016) and CARE (Carroll et. al., 2020) principles. This report ties the workshop breakout discussion content into a comprehensive outline for geoscience research data commons requirements.

2.1 Report Structure

This report begins with a brief overview of the workshop scope, goals, and structure. We then provide a set of key observations and associated recommendations that derive from the workshop discussions which are augmented with the following appendices that provide additional supplementary information related to the workshop:

- Appendix I - Workshop participants and steering committee members
- Appendix II - Workshop agenda
- Appendix III - Plenary Sessions - Detailed Agendas, Titles and Bios
- Appendix IV - Breakout session discussion questions by workshop theme

2.2 Workshop Background and Goals

The workshop concept was motivated by an internal initiative at NSF NCAR to review the existing architecture of its data services portfolio, and investigate if that architecture is structured to best support the emerging needs of modern data science research workflows and community data and software sharing expectations. NSF NCAR's existing data services architecture is structured as a set of siloed, domain specific repositories and data sharing websites that have organically evolved over time and are designed to support the 'download and analyze' model that has traditionally been used by their targeted domain science research communities. While this architecture has worked well for many years, it creates challenges for researchers who want to: 1) analyze large data volumes (10s to 1000s of TBs in size), 2) work with heterogeneous datasets (e.g., model outputs and observations), 3) publish datasets to support community open science expectations, and 4) explore convergent, interdisciplinary research questions. Additionally this architecture can be confusing and challenging to use, thus limiting who can leverage NSF NCAR's data assets in modern data science research and educational applications to only those with expert knowledge of these systems. One goal of NSF NCAR's initiative to examine and restructure its data services portfolio is to simplify its usability and broaden participation in data science research and education.

The workshop brought together a diverse group of over 70 participants to provide a variety of perspectives on community needs for a geoscience research data commons environment. Stakeholder communities represented at the workshop included geoscience researchers and students (weighted towards the Atmospheric, Hydrologic, Geospace and Oceanic sciences) from US Universities (including Minority Serving Institutions (MSIs) and Historically Black Colleges and Universities (HBCUs)), technology experts, representatives from open source communities of practice such as Pangeo [5], private sector organizations, and data repository personnel. Workshop participants were recruited through steering committee member recommendations and announcements to solicit participation from the broader community through a number of venues, including the Minority Serving Cyberinfrastructure Consortium, UCAR Education, Engagement & Early Career Development Program, UCAR Currents

[5] https://pangeo.io/

newsletter, Unidata Users Committee, the UCAR booth and American Meteorological Society (AMS) data help desk during the AMS 2024 annual meeting in Baltimore, MD, the Coupling, Energetics, and Dynamics of Atmospheric Regions (CEDAR) listserv, and the Space Physics and Aeronomy (SPA) newsletter in geospace science.

The overall goal of the workshop was to ask participants to discuss topics that would inform a workshop report which describes community recommendations for geoscience research data commons needs from a broad range of perspectives. Additionally, we wanted to build a welcoming and inclusive community through the workshop that will provide the foundation for continued collaborative activities that result from the workshop. One such activity will involve participation in data commons user focus groups that will guide modernization efforts of NSF NCAR's data services offerings.

2.3 Workshop Structure

The two day workshop was organized with three predefined plenary and breakout discussion session themes, that included:

1. Big Ideas and Existing Solutions (for data services and data analytics capabilities)
2. Data Commons Analytics Research and Educational User Needs
3. Data Producer Needs for Data Curation and Consulting Services

For each of these themes there was a plenary session that included 5 topical talks to promote ideas and stimulate discussion in breakout sessions that followed. The breakout sessions included 7 separate breakout groups that workshop participants were randomly assigned to, and each of these groups discussed identical theme questions for these first 3 workshop themes. One exception to the random breakout group assignment occurred during the "Data Commons Analytics Research and Educational User Needs" session, where students and early career professionals were assigned to one breakout group to promote ease of discussion and networking amongst this cohort.

A 4th breakout session took place during the last day of the workshop. For this session, workshop participants were given the opportunity to suggest and vote on breakout group themes that they felt would be of value for further discussion. Throughout the workshop, participants were asked to provide breakout topic suggestions through an online platform, which they could then up or downvote topics as the workshop progressed. Workshop organizers reviewed the suggested topics and provided a final opportunity for voting during the final morning of the workshop. From here seven separate breakout group themes were chosen and workshop participants could self-select which breakout session they attended. These final breakout group topics included:

1. Education and Outreach and community building - what approaches work for different communities? What has been successful so far?
2. What collaborations can we initiate at this meeting to work on concrete next steps identified by the workshop, including timeline?
3. Policies, including data accession and deaccession in data repositories (e.g., data retention and data deletion policies). How to define and what criteria can be used to

evaluate?

4. Expanding accessibility and standardization of processes within the framework of limited funding -what are actionable tasks that can be community driven with little to no financial investment?
5. What are the barriers to adopting cloud computing and High Performance Computing?
6. Where does equity in data discoverability and access, and suitability for research purpose fit into a data commons? I.e., with open access, we might still not have equitable discovery.
7. What does your vision for a data commons look like now? Gap analysis?

The following sections of this report provide key takeaways and further details from the discussions held in each of the 4 breakout sessions. We then provide recommendations associated with key discussion topics.

2.4 Why Data Commons?

This workshop was structured around the concept of “data commons” to build on national and international expertise in the development of data infrastructure, tools, and services. The term “commons” itself derives from research in economics and environmental science on “common pool resources,” such as fishing grounds, forests, and grazing lands. Such “commons” have been extensively studied to understand how they are developed, made productive, and sustained.[6] Applying this concept to digital resources and scholarship, Grossman describes a data commons as follows:

> "We view a data commons as a software platform that co-locates: 1) data, 2) cloud-based computing infrastructure, and 3) software applications, tools and services to create a governed resource for managing, analyzing, and sharing data with a community. Briefly, a data commons is a cloud-based software platform with a governance structure that allows a community to manage, analyze and share its data."[7]

Grossman’s description assumes a cloud-based technology, but cloud computing and storage infrastructures are only one possibility as the underpinning for a data commons. As Grossman’s definition notes, a community orientation and solid governance structure are essential for the operation of a data commons. This is a characteristic of effective commons of any kind:

> “Research on commons governance posits that sustainable, successful commons have clear boundaries, complex governance rules, and active management (Ostrom, 1990).”[8]

---

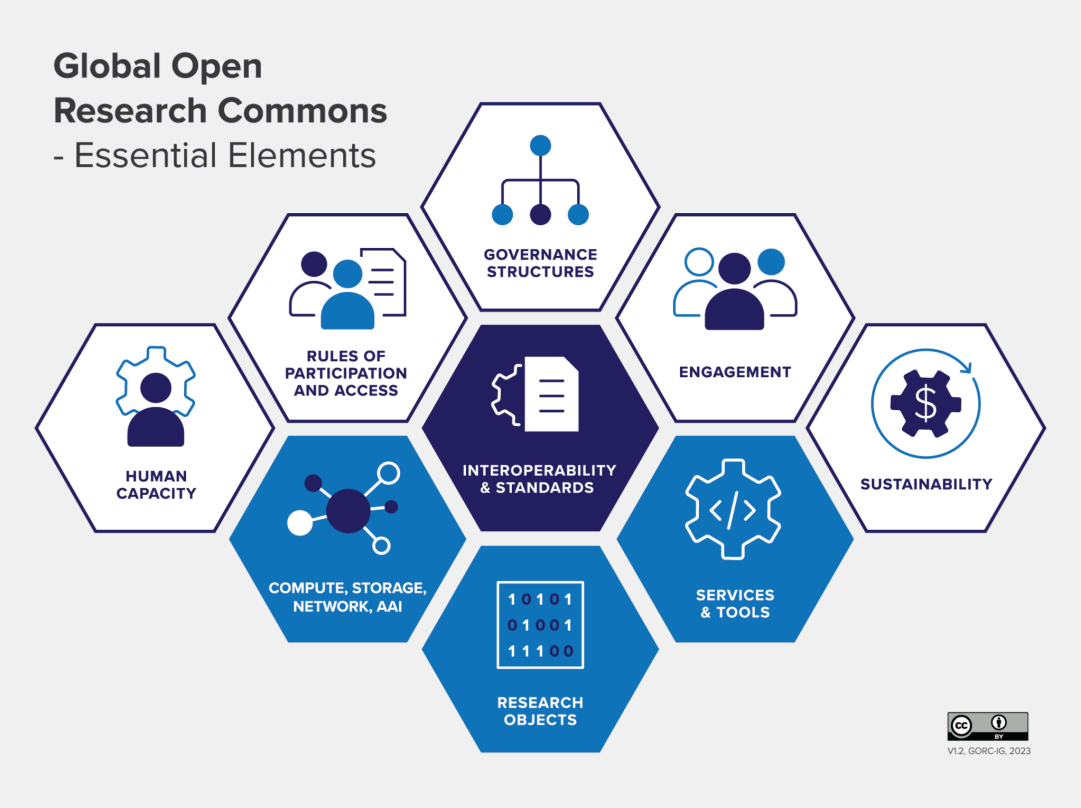
[6] Fisher, J. B., & Fortmann, L. (2010). Governing the data commons: Policy, practice, and the advancement of science. *Information & Management* 47(4): 237-245. https://doi.org/10.1016/j.im.2010.04.001

[7] Pg. 1, from: Grossman, R.L. (2023). Ten lessons for data sharing with a data commons. *Scientific Data* 10, 120. https://doi.org/10.1038/s41597-023-02029-x

[8] Pg. 1757, from: Eschenfelder, K. R., & Johnson, A. (2014). Managing the data commons: Controlled sharing of scholarly data. *Journal of the Association for Information Science and Technology* 65(9): 1757-1774. https://doi.org/10.1002/asi.23086

This multifaceted nature of a commons is demonstrated by Figure 1, the essential elements of a research commons, which was developed by an international group of experts, based on detailed analysis of 13 large scale research-focused commons initiatives.[9] The elements shown in Figure 1 encompass the components of a commons to support data-driven science.

Figure 1 - Global Open Research Commons - Typology of Essential Elements



These conceptions of a data commons from Grossman and Figure 1 were presented at the beginning of our workshop, to set the stage for a multifaceted discussion of technical needs, governance challenges, and community engagement requirements, amongst other topics.

3. Key Outcomes from Workshop Discussions

This section presents key outcomes from the workshop discussions. Given that many of these topics were discussed across multiple sessions, content from all breakout sessions were synthesized together to inform these outcomes. This section is organized around the

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

[9] Jones, S., Leggott, M., Lopez Albacete, J., Pascu, C., Payne, K., Schouppe, M., Treloar, A., & Global Open Research Commons IG. (2023). *GORC IG: Typology and Definitions (1.01)*. Zenodo. https://doi.org/10.15497/RDA00087

components of the Global Open Research Commons model shown in Figure 1, as this provides a robust framework for structuring the topics that were discussed during our workshop.

3.1 Governance Structures

As noted in the introduction section where the term “data commons” was discussed, governance is a key component of any commons. In the case of a data commons, governance is needed to ensure that the collections, tools, and services have proper oversight, direction, and operation. Governance should involve both the data commons provider and participation from the relevant science communities. It will be important to find a balance between democratizing data governance and maintaining data integrity and fidelity. A governance council should be created to steer the directions of the data commons.

A number of specific topics were discussed in the workshop that will need to be addressed via governance policies. First, policies for publicly sharing datasets need to be modernized. Governance policies for data uploaded to a commons will need to deal with quality control, including dealing with bad actors and correcting unintentional mistakes made by data providers. On evaluating data submissions, the comment was made as follows: “Trust, but verify!” This is particularly important as a commons is opened to more people and data types. For example, sovereignty and privacy are concerns for social and human-subjects data. A commons might increase trustworthiness for citizen-produced data by having workflows, community forums, and FAQs to address questions. Governance structures should also respect data control and sovereignty, following the CARE Principles for Indigenous Data Governance[10] when appropriate. Overall, it will be necessary to determine “sweet spot” policies that describe the level of services provided for all data.

Second, software governance is as important as data governance. If a data commons provides access to community developed software and/or computing platforms, it will be important to have governance policies related to software documentation, modification, and reuse.

Third, there is a need for data and software persistence and permanence guidelines. For both, persistent identifiers like DOIs help with permanence. For software, sharing code exposes the underlying algorithms, which can help ensure reproducibility of a software workflow, even if it is ported to a new code base. Data deletion policies will be needed to prioritize which data are stored and maintained for the long term, rather than trying to keep everything for perpetuity.

But how are data and software retention and deletion policies defined, and who defines them? Many key questions were discussed about how to measure the impact of data and software. How to evaluate if a dataset will be impactful to the community? How many users are there now, and will there be in the future? Is there a sunsetting strategy for selected data types? What data needs to be preserved and shared to meet evolving requirements of publishers, funders (e.g., open science expectations), and government based regulations (e.g., federal open data statutes)? How does the impact of data change as a function of time? We need community

[10] https://www.gida-global.org/care

accepted evaluations, but with regards to purge data/software, is it all about the numbers? What qualitative factors should also be considered?

Governance policies are also needed to clarify types and levels of service for data contributed to a commons. For example, there may be differing levels of curated data starting with "minimally/self curated and shared/published with no DOI" that will age off in a short time period by default. These could be published through "Github-like" pull requests. The other end of the range is datasets that are "fully curated with a DOI in a trusted repository." Governance is needed to determine and specify how a dataset is elevated from minimally to fully curated data.

Finally, what are the timelines for serving different versions of a given dataset? There is a need to balance quality control with immediate data availability. The incentives for data sharing and appropriate roles of data embargoes are complicated, and should be specified.

RECOMMENDATIONS

- ***A community informed governance strategy is needed to develop policies to structure processes for data deposit, documentation, access, retention, deletion, and preservation.***

3.2 Engagement

Workshop discussions made clear the importance of community building around data commons. Two-way communication is required between the user community and developers. Such communication should be emphasized, incentivized, and encouraged. This can promote new technologies and solutions. Workshop plenary presentations and breakout discussions also emphasized that engagement must go beyond R1 research institutions, and include targeted and ongoing engagement with potential users from non-R1 universities, Minority Serving Institutions (MSIs), Historically Black Colleges and Universities (HBCUs), and other institutions that have limited resources.

Specific types of user engagement were discussed. On-boarding, in particular, is key for the success of a data commons. Early career and later career individuals have different on-boarding needs. Further, users may be far more diverse than the traditional scientific community, particularly when non-scientists and cross-disciplinary scientists are being served, making equitable and inclusive on-boarding approaches even more important. On-boarding should involve making the benefits immediate and simple via training and skill development. Specific approaches to on-boarding include focused training, hackathon weeks, participation in analysis working groups, interacting with communities directly, providing galleries of worked examples and use cases, and extending known tools to connect with your data commons. Other steps that enable easier on-boarding of users include: Providing a means to analyze and visualize the datasets, e.g. via sample Jupyter notebooks, providing both detailed and simple plain-language documentation, and making data and software versions easily citable by providing DOIs and recommended citations.

An emphasis on education and training toward best practices helps to ensure fair and equitable access for all users, including users from MSIs, HBCUs, and other traditionally underrepresented users (e.g. community colleges or non-academic community groups). Targeted education and training is particularly important for these groups, as they may lack resources to develop or host data-focused training themselves. Further discussion touched on the potential for pathfinders, that is, platform capabilities where subcommunities can form and collaborate. This was also discussed via the idea of giving people the tools to build their own data commons. There will never be one single data commons that supports all science communities and needs, but there would be value in commons' providing capabilities that allow communities to build and coalesce around particular data and tools.

On-boarding was recognized as a major challenge. Education is necessary to fill curriculum gaps, particularly in techniques and terminology involved in software programming, but also in more mundane tasks such as properly phrasing data search requests, which emerging AI tools may also be able to support. How do we effectively educate a workforce that knows how to take advantage of what a data commons might offer? What's the ideal breakdown of that workforce, e.g. among those who can use existing Python code, those who can adapt existing code to fit a use case, and those who actually develop new tools / move the needle? Finally, how to instill a lifelong practice of learning and sharing new things?

RECOMMENDATION

- ***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.***
- ***Engagement strategies for a data commons must include outreach to all types of users, especially including traditionally underrepresented communities from non-R1 universities, MSIs, HBCUs, and other institutions that may have limited resources.***

3.3 Rules of Participation and Access

This section highlights the need for clear rules of participation and access regarding the resources and services made available via a data commons. These rules should focus on the needs of users, whoever they might be, with the goal of expanding data access beyond existing user communities. Important questions here are: Who is the community being served by a data commons? Will a data commons be defined by the end-goal community or the type of data stored? FAIR data access potentially opens doors to broad access by more user groups than in the past, but broad access must also come with clear rules and requirements for data deposit and for access and use of any associated computational tools.

There are a full spectrum of different users, use cases, and requirements that need to be accounted for (e.g. academic researchers, public sector scientists, social scientists, private sector, non-scientists, etc.). But the scale of the user community must be balanced with issues

such as scalability with limited funding. There are large implications for domain-specific developments, such as API support for variable data formats and analysis tools.

There is interest in building partnerships with private sector companies. This idea comes with a number of important considerations regarding rules of participation and access:

- It will be necessary to define clear intentions in writing prior to working with private companies.
- Transparency goes both ways. Companies should be clear on how much money they are making off of common data.
- Would it be necessary, beneficial, or hurtful to have agreements that prevented profits from research data?
- Regarding rules of access, should there be differences in non-profit sector (.edu) vs private sector regarding the cost of access and use? Should .edu email addresses get more free tier resources?
- What partnership models are sustainable? Can private sector companies pay for additional features that also benefit public sector researchers? Is it a private sector role to serve as the intermediaries between public sector data commons and private sector data consumers, for example, by building analytics platforms in the commercial cloud for downstream private sector data users?
- What discussions are needed between different agencies to get similar agreements out of partnerships with commercial cloud providers?

RECOMMENDATIONS

- ***A data commons must have well-defined rules, policies and licenses describing what users are able to do with its data, software, and services. These rules and policies should define boundaries around usage, including deposit, storage, and access of data, and access and usage of software or computing services.***
    - ***Guidance should be provided to data depositors on what data they need to preserve and share and for how long (e.g., raw observations, statistically processed observations, model outputs).***
    - ***Tiered levels of value added services could be made available and provisioned through an equitable allocations process, similar to the NSF ACCESS[11] strategy.***
        - ***Value added services, such as hosted computational notebook resources, could especially open up access to data science research for traditionally underrepresented communities from non-R1 universities, MSIs, HBCUs, and other institutions that may have limited resources.***
    - ***Policies for the use and management of resources (data & infrastructure) need to be clearly stated to user communities, and technical capabilities should ideally be developed to enforce such policies.***

[11] https://access-ci.org/

3.4 Human Capacity

This section encompasses the human expertise and support needed to ensure effective operation and use of a data commons. From workshop discussions, it is clear that there are unmet resource needs for infrastructure support, but many of the challenges researchers face may have existing solutions (e.g. on how to deal with data format challenges). Researchers need skilled programmers, however, to help solve these problems, along with other necessary issues, such as cybersecurity concerns with open source building blocks. There is a lack of training on data and computing fundamentals. Further, there is a lot of "unseen/hidden" data curation work already happening as well within research teams. Funding agency grants are generally not targeted to supporting these kinds of work, and scientists do not have the time to learn all of the required skills. Universities also often do not support operational data infrastructure and associated staff.

Thus there is a need for experts of various types to support researchers' data needs, including programmers, data curators, and facilitators who can connect people to the right experts. It is important to value these data curators and software communities who create and maintain useful data sets and software packages.

Finally, the different stakeholders involved in the operation of a Data Commons (data providers, service providers, DC admins, …) require well defined operational methodologies (coordination, communication,...) in order to support successful operations. This effort and dedicated resources is not to be underestimated.

RECOMMENDATIONS

- ***There is a strong desire for sponsor-supported centralized data curation resources, similar to the NASA Data Repositories dataset assignment model [12]. In this model, the funder provides a designated repository where datasets that result from funded projects can be deposited along with a chosen level of curation support. Such a service will need well defined expectations and scope of service. The benefits of such a service include:***
    - ***Data providers could focus their efforts on data collection and analysis.***
    - ***Data management planning and implementation would be more straightforward***
    - ***Oversight in the proposal review process would be easier***
    - ***Support for data management & preservation could begin at the proposal stage***

3.5 Sustainability

Geoscience data commons providers must be trustable and sustainable entities. Sustainability of a data commons includes durability of the data, associated tools, support staff, plus any specialized services that are provided on top of the core infrastructure capabilities. Workshop

[12] https://pub.earthdata.nasa.gov/data_publication_guidelines

participants discussed how the traditional 3-5 year research grant funding model is not the right one for providing persistent community services like data commons. How do we address sustainability and ensure services are designed in ways that allow them to continue to persist or migrate elsewhere when, for example, grants expire?

Participants also noted that there is a lack of insight into how much data services actually cost. Invisible labor is happening and not funded appropriately, e.g. data curation, software development, and user support. The cost differences between cloud vs. on-premise data services are also often opaque, including both current operational costs and future costs of sustaining such resources. There is a need for more transparency regarding the resources required to build, operate, and sustain a data commons.

Sustainability of an infrastructure involves a range of socio-technical considerations. First, where are the resources going to come from and how will they ideally/equitably be allocated? Second, at a community level, how do you incentivize contributing data or software to a group effort like a data commons? There is a need to identify and specify funding for each stage of the data life cycle. Sustainable funding models might involve considering costs associated with different levels of service. For example, depending on user needs, there might be different costs associated with different services, such as data deposit, access, or analytics tools. However, it will be essential to consider equity issues if moving to some sort of premium cost-recovery model. Including data preservation and reuse funding as part of a scientific research project's funding proposal is another way to provide funding for data infrastructure.

Sustainability is a core part of being considered a "trusted repository," as it is a foundational requirement of the CoreTrustSeal Data Repository Certification[13], as well as other similar certifications. Users need assurances that the data commons will persist. Persistence and sustainability are not just about data storage. They also involve constant attention to data curation, service continuity, and preventing data decay.

RECOMMENDATIONS

- ***A data commons needs a business model to support sustained infrastructure, including the people, storage, computing resources, and protocols involved.***
- ***There must be a long and predictable funding lifecycle for a data commons to be sustainable and meet community expectations for being a trusted service.***

3.6 Interoperability and Standards

Data portability and interoperability require compliance with standards. Standards are also necessary for any data commons to access data from multiple sources, e.g. through the use of software agnostic APIs that adhere to community conventions.

Standardization can be a significant challenge, however. For example, how should a data commons ensure standardization of data formats and metadata for small and large datasets?

[13] https://www.coretrustseal.org/

Some communities have existing data formats/standards that have coalesced, such as the Climate and Forecast Metadata Conventions [14], but there is a lack of data and metadata standards in some communities. A data commons would need a general metadata schema, such as the DataCite Metadata Schema [15], plus domain specific schemas or extensions to the general schema. To simplify metadata creation, it will be important to abstract out metadata details from researchers that need to create metadata, like a turbotax [16] for metadata. Large language models (LLMs) can assist with metadata generation by pulling information out of abstracts or other documentation. It would be helpful to organize workshops and analysis working groups involving experts to help standardize specifications for metadata, and perform education and outreach activities.

It is not always clear what kind of metadata is actually needed to drive scientific workflows. A data commons should not set the bar too high for data producers when it comes to adhering to standards, as they may not have requisite expertise or experience with standards. There are many domain-specific metadata standards, and it is not reasonable to expect that scientists are aware or understand all of them. Tools and consulting support needs to be provided to abstract out the details of metadata development for data depositors.

These challenges to publicly sharing datasets become more acute for uncommon data types or formats, e.g. specifying data and metadata standards and developing software stacks for data processing. Likewise, open questions remain about how a data commons should handle legacy or dying data formats. Will it be necessary to keep providing software handlers for legacy formats or convert a legacy dataset to a modern format? These would require software and data maintenance. Instrumentation is also a potential challenge, as many instruments produce data in proprietary formats. Workshop participants expressed a hope that new cloud-based formats such as Zarr [17] and Apache Parquet [18] can solve some of these problems as they have common tooling, and are language agnostic and self describing. But it is difficult to predict technology changes that might drive the use of new or different data and metadata formats.

RECOMMENDATIONS

- ***Data and metadata standards and best practices on using them need to be determined by domain science communities in conjunction with data commons personnel, followed by work toward crosswalks and interoperability tools across those community solutions.***
- ***Tools and consulting support needs to be provided to abstract out the details of metadata development for data depositors.***
- ***A data commons should provide standards-based APIs to abstract out complexities and enable simplified data and metadata access.***

[14] https://cfconventions.org/
[15] https://schema.datacite.org/
[16] https://turbotax.intuit.com/
[17] https://zarr.dev/
[18] https://parquet.apache.org/

- ***Data Commons should strive to maximize interoperability with peer systems at the national and international level (e.g., NASA, NOAA, NSF, ESA, Copernicus, etc..).***
- ***Broad and active participation on standardization forums (e.g., Open Geospatial Consortium) at the national and international level will facilitate wider interoperability and synergies with data commons systems worldwide (international and cross-domain).***

3.7 Compute, Storage, Network, AI

A data commons should be built as an open platform that includes data from disparate sources, and has associated scalable compute resources. This will ensure that data is findable as well as easily processed and analyzed. Barriers exist regarding the best tools for data access and storage, authorization/authentication, and other key components (e.g. streaming data access via https). Support for data proximate computation is also lacking or minimal at present for many communities.

There are good models for how a data commons technical infrastructure can come together. European agencies have been using a federation concept based on a distributed cloud model to help align resources and define centers of excellence. In other cases, workshop participants that were using data analytics platforms such as SciServer [19] and CryoCloud [20] were generally happy with these services and struggled to identify gaps. There is also a need to advertise existing computing capabilities that are available to the science community, such as the NSF Access [21], OS Pool [22], and NAIRR [23] resources.

Ideally, data should be performantly accessible from all computing platforms, including high performance computers (HPC), research clusters, the commercial cloud, and laptop-based streaming access. In this vision, all computing services are “data proximate” regardless of where the authoritative data repository is located, such as with the Open Science Data Federation (OSDF)[24] or Tigris data service[25]. This model will allow any stakeholder to be creative and build their own data analytics tools on top of the data commons however they would like to do so, without the need for data replication.

Workshop participants noted, however, that commercial cloud-based services also present risks. It is important to acknowledge that commercial cloud storage might not always be sustainable, and can be ephemeral. In addition, a “one solution, one cloud” approach is impractical. There will need to be a multi-faceted approach that uses commercial cloud where appropriate without assuming that commercial cloud services will solve all of the needs of the science community. With this in mind, workshop participants were interested in exploring how the science

[19] https://www.sciserver.org/
[20] https://cryointhecloud.com/
[21] https://access-ci.org/
[22] https://osg-htc.org/services/open_science_pool.html
[23] https://nairrpilot.org/
[24] https://osg-htc.org/services/osdf.html
[25] https://tigrisdata.com

community might partner up with commercial cloud providers and other technology providers in non-academic sectors.

There were also many discussions of ongoing developments in artificial intelligence and machine learning. There was interest in leveraging natural language models and large language models (LLMs) for UI/UX and data discoverability, detecting anomalous data values, and for extracting data values from the data commons in response to specific user research questions. It was acknowledged that findability and searchability of datasets is difficult. Can large language models be used to improve search and discovery, and can they help develop crosswalks across metadata standards?

Beyond using LLMs for search and data extraction, there was discussion of machine learning tools that were more targeted toward geoscience research. For example, there was interest in having a data commons host or provide federated access to lightweight, pre-trained, and tunable machine learning models, such as Microsoft's Aurora Foundational Model of the Atmosphere [26]. In addition, there was interest in developing machine learning software with the capability to ingest earth science datasets, for example how tensorflow [27] can accept NetCDF-4/HDF-5 chunks.

RECOMMENDATIONS

- ***Data storage and access solutions, whether based on cloud services or other technical approaches, must ensure that both short and long term data preservation and sharing are supported.***
- ***Storage, access, and analysis services must also scale appropriately for different volumes of data, and ideally storage will be performantly accessible from many types of compute services to avoid the need for replication of datasets.***
- ***Leverage the capabilities provided by Large Language Models to:***
    - ***Simplify data discovery and access***
    - ***Automate metadata extraction***
- ***Provide access to community developed lightweight, pre-trained, and tunable machine learning models.***
- ***Data commons developers must envisage technical capacity for increasing user demand and account for cybersecurity considerations.***
- ***Data commons should be architected with a flexible cyberinfrastructure that will allow for the migration of data & services between differing platforms and infrastructures.***

3.8 Research Objects

As should be apparent from the previous sections, in the geosciences there are many different data types being produced by many different people and scientific facilities. The research

[26] https://doi.org/10.48550/arXiv.2405.13063
[27] https://www.tensorflow.org/

objects being generated, stored, and distributed are thus highly variable in many characteristics. Metadata for these objects likewise varies significantly. Data commons must be designed to either handle this variability, or to reduce it via technical, human, and/or governance means. The challenge here is in helping users to scope and scale diverse data collections down to useful information.

Another key consideration for a data commons is whether the commons will be providing access to raw vs processed data. “Raw data” is itself a somewhat contentious concept, as different people will interpret it differently. But workshop participants noted that raw data can be difficult or impossible to use, and that some degree of data processing is necessary to enable data use. Very processed data can also be difficult to use as well, however, if the processing is too bespoke and limits further investigation. A middle ground is desirable, where processing results in understandable and consistent data, while room is left to enable innovation for end users.

Documentation and education by the data providers about their data is important to build data trust, and help users to interpret data and their context. As noted above, features like community forums, blogs, links to papers, help desks, and discussion channels could be useful data commons features for this purpose.

Open questions for a data commons surround the extent to which data updates may happen, and whether derived datasets will be in scope for collection. In other words, when a data commons provides access to data, how will changes or derived versions of those datasets be handled? Workshop participants noted that a data commons needs to store community generated derived data products, and document linkages to parent datasets. Metadata curation and data maintenance is important as data are manipulated, updated, or reprocessed. The importance of data provenance cannot be underestimated here, but data and metadata provenance tracking should be automated whenever possible. Provenance information is difficult to generate manually, particularly after the fact.

Additional questions were raised but not discussed in detail about what level of metadata provides optimal value? Can we minimize the requirements for metadata and modernize our curation processes to maximize value for the level of effort required? What is the minimum level of metadata required that presents the lowest burden for entry? How should those requirements change for each dataset, potentially taking into consideration the expected impact (or lack thereof) for the dataset? These are important questions that could be investigated further in future workshops.

RECOMMENDATIONS

- ***Minimum requirements and the purpose for dataset metadata should be well described (e.g., support search, provenance and modern data science needs).***
    - ***Metadata generation and provenance tracking should be automated where possible.***

    - ***The metadata requirements should vary depending on the properties and expected impact of the dataset.***
    - ***Persistent Identifiers should be assigned to dataset landing pages and possibly individual data objects where appropriate***
- ***A mechanism should be provided that allows community feedback for datasets and their metadata (e.g., similar to GitHub issues).***

3.9 Services and Tools

Workshop participants discussed many different services and tools that a data commons might provide. These services should be community-developed and informed, should address science needs, and build on existing practices and resources where possible. Specialization and community developed designs can provide use case successes (such as Xarray [28], Climate Data Operators [29], NetCDF Operators [30]) that can be used as a model for data accessibility, etc.

Adoption of any services or tools will be enabled by great user experience (UI/UX) and design. As a starting point, services are needed to enable data discovery, access, and training to use data. User-friendly data search interfaces are essential, along with a website or knowledge space where people can go to look for help. An ability to discuss datasets in the same place as the data is stored would be really useful. Connecting and augmenting data was also a theme. Datasets should provide a place for users to upload their own use cases tied to that dataset. A data commons could integrate some of what GitHub offers, such as the ability to “star” a dataset, and to provide and maintain a user forum. If these are implemented, however, it is important to adopt a code-of-conduct so anyone can feel safe asking a question. Connecting those discussions and related resources to the dataset is expected to improve the reusability of the dataset over time.

Workshop participants acknowledged that the pace of technical change is overwhelming. How do scientists deal with this, and how will a data commons deal with this pace? Scientists do not have the bandwidth to keep up. There was discussion of containerizing environments, to freeze them in certain configurations that are known to be useful or need to be preserved for some specific reason. Such an advance would improve the reproducibility of a given work if it references a given version of that software container. There was also discussion on how private sector partnerships could assist public sector entities in keeping pace with the rapidly changing technology landscape.

Jupyter notebooks were called out a lot as a great tool for supporting research & educational use cases, as there is an ease of use for Jupyter-based resources, and good options for sharing notebooks. JupyterHub is a powerful tool, but is still used locally for the majority of use cases. Resources are desired for university professors to get access to JupyterHubs hosted by a data commons, along with allocations of computing resources. Such services could allow users to

[28] https://xarray.dev/
[29] https://code.mpimet.mpg.de/projects/cdo
[30] https://nco.sourceforge.net/

visualize and stream data without downloading the data. Additionally, a data commons should enable communities to describe, test and share their workflows for higher level post-processing of data. Project Pythia[31] is a good example, by providing notebooks with embedded expertise to expose data and software to new users.

In general, workshop participants were interested in technical solutions for supporting robust metadata generation, to empower scientists without the burden of requiring them to intimately know many metadata standards. Software and intuitively designed user interfaces will help solve these challenges.

RECOMMENDATIONS

- ***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.***
- ***Integrate open source community developed software solutions into data analysis and access tools.***
- ***Pursue public/private sector collaboration and partnerships to keep up with the technological pace of change and deliver modern/current services and tools.***

3.10 Attendee-Selected Discussion Topics

This section provides brief overviews of the topics and discussions chosen by the workshop participants for the fourth and final breakout session. We note here points discussed in these sections that are complementary to the topics outlined above, and serve to highlight key concerns of the workshop attendees.

Vision for a data commons:

- All types of data common to a community should work in the commons. But we can't build one data commons to rule them all.
- It is desirable to be community focused, while reducing confusion about where to store data, what methods of access are available, and what level of data curation is needed.

Collaborations required for concrete next steps:

- International collaboration opportunities, e.g. European Commission, Open Geospatial Consortium, NASA, European Space Agency
- How to keep the momentum from this workshop going? Ideas include: Annual meeting, online communication channels, engaging with other Data Commons efforts, attending the yearly NASA Data Repository workshop. and engaging with university faculty and students.

Equity in discoverability and fit for purpose:

[31] https://projectpythia.org/

- Build ties between data and infrastructure providers and users.
- Provide financial support for educators/leaders/scientists to be trained to train others.

Education and outreach and community building:

- Existing Successful Approaches
    - Synchronous training: summer schools and workshops, e.g. MetPy Mondays.
    - Reproducible Asynchronous Workflows: e.g. Pythia Cookbooks, Pangeo Gallery
- Challenges - What can be done better?
    - Training material targeting various levels of users, including future data curators
    - Freemium access to data & compute resources.
    - Keeping curricula up to date

Data accession and deaccession in data repositories:

- Data retention policies vary. It would be helpful to do a survey of these policies.
- Many data retention criteria exist: duplication, past use, scale of interest, can it be reproduced (for model output), timeline of utility (e.g. a couple of years after the paper comes out; long enough to create smaller derived products); uniqueness, role as reference points for later work. “What is the risk if lost?”

Expanding accessibility and standardization within the framework of limited funding

- A small initial investment upfront in documentation and examples will grow and save time in the long run, such as code snippets, links to publications, and examples.
- Community involvement, crowdsourcing, incentivizing contributions are all important.

Barriers to adopting cloud computing and high performance computing (HPC)

- Agencies need to lead and ease onramps for groups that may be reluctant to adopt cloud computing.
- The typical 3-5 year funding model still makes it difficult to move to cloud computing instead of locally-housed computing hardware.

3.11 Overarching Comments

This section provides a brief overview of recurring topics that came up during both plenary and breakout session discussion throughout the workshop, but did not necessarily fit into one of the essential elements of a research commons model.

Plenty of gaps exist in data service needs. While many cylinders of excellence exist to provide data services, workshop participants noted that numerous gaps still exist in the repository ecosystem for supporting publisher data sharing requirements, and this leads to many questions including, “Where should my data be archived?”, “What capabilities do data repositories provide?”, “Who pays for long-term data curation?”, and “How do we avoid silos?”. Additionally participants discussed the role of domain-specific repositories vs generalist repositories and a future vision for separating storage from discovery, access, and analytics tools.

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. Ultimately, the value in data is in its use and application, which led to the questions, "How does one identify potential users and communities who would be interested in using a dataset?", and "How can we cater to different users' needs and requirements?". The user base for a data commons should be broad and diverse. This requires equitable and inclusive governance and technology, as well as education and training opportunities. Developing governance goals and engagement strategies for specific user communities such as MSIs and other underrepresented institutions should be done in concert with these communities themselves. Finally it was noted that developers of a data commons should define a North Star to guide them in their development. For example, "What are we trying to optimize with the creation of a data commons? Ease of Use? Cost effectiveness? Research Innovation? Others? -Pick two."

What is the common denominator for a data commons? There was much discussion in the workshop about "what the common denominator for a data commons should be?", following a plenary presentation on this topic. 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). Also, there are many common technical needs for all use cases (e.g., storage, backups, etc), but conversely there are domain-specific aspects that need to be supported (e.g. data formats, metadata, how the data are used). Participants emphasized the need for a data commons that provides the foundation to build bespoke solutions upon, similar to the collaborative development platform concept (e.g., GitHub) that provides a common platform on which different communities can build and grow upon. This will allow communities to innovate and develop their own bespoke solutions on top of a data commons infrastructure vs having the commons build narrowly focussed, bespoke solutions to serve niche communities. Finally, most participants agreed that a data commons will ideally abstract out complexities, such as unique data formats, and provide simplified data search and access capabilities.

Culture change is needed along with (perhaps new) business models. The topic of culture change and incentives to share data and code repeatedly came up throughout the workshop. Participants agreed that incentive structures in academia need to change to reward activities that support data curation and open science needs vs hoarding of resources by individual groups to produce as many peer reviewed publications as possible. Finally, presented with the challenge of static funding environments and an accelerating pace of technological change, workshop participants concurred that there is an emerging need for more public/private sector partnerships to give researchers access to the latest technological capabilities.

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 [32]) to instill trust in their downstream user communities and stakeholders. Additionally, participants emphasized that sustained, predictable funding to make impactful data and software reusable by others, which include the services provided by a data commons environment, will be needed to instill long-term community trust in such a resource.

4. Conclusions and Next Steps

A wide range of community needs and the path forward to support those needs through a data commons resource were discussed during the workshop. The recommendations outlined in this report are intended to provide a foundational framework for those developing data commons services to build upon. In addition to these recommendations, a number of overarching themes and outstanding questions that need further exploration arose. Many connections were made between those who participated in the workshop, and we plan to continue engagement amongst workshop participants through our common email list, engage the broader community at relevant domain conferences such as the AGU and AMS annual meetings, and by soliciting workshop participants to participate in follow-on user focus groups that will inform the development of NSF NCAR's geoscience research data commons service.

Appendices

Appendix I - Workshop participants and steering committee members

Participants: Agbeli Ameko (NSF NCAR), Michael Bell (Colorado State University), Asti Bhatt (SRI International), Brian Bockelman (Morgridge Institute for Research), Bernadette Boscoe (Southern Oregon University), Asaya Bulgin (Elizabeth City State University), Prashanth BusiReddyGari (University of North Carolina at Pembroke), Katherine Cariglia (MIT Haystack Observatory), Deepak Cherian (Earthmover PBC), John Clyne (NSF NCAR), MaKenna Collins (Jackson State University), Scott Collis (Argonne National Laboratory), Riley Conroy (NSF NCAR), Ian Cornejo (University of Wisconsin - Madison), Nick Cote (NSF NCAR), Tom Cram (NSF NCAR), Chris Crosby (OpenTopography / EarthScope), Alisdair Davey (National Solar Observatory), Steve Diggs (University of California), Mohamed Elbakary (Elizabeth City State University), Orhan Eroglu (NSF NCAR), Ana Espinoza (NSF Unidata), Clark Evans (University of Wisconsin - Milwaukee), David John Gagne (NSF NCAR), Jiwon Gim (NSF NCAR), Kevin Goebbert (Valparaiso University), Max Grover (Argonne National Laboratory), Joseph Gum (NSF NCAR), Thomas Haine (Johns Hopkins University), Thomas Hauser (NSF NCAR), Nils Hempelmann (Open Geospatial Consortium), Nathan Hook (NSF NCAR), Brenda Javornik

[32] https://www.gida-global.org/care

(NSF NCAR), Chenyue Jiao (University of Illinois Urbana Champaign), Miguel Jimenez-Urias (OPeNDAP), Hailey Johnson (NSF Unidata), Jhordanne Jones (UCAR), Nakul Karle (Howard University), Eric Kihn (NOAA NCEI), Teagan King (NSF NCAR), Ilene Locker Carpenter (HPE), Scot Loehrer (NSF NCAR), Angel Lopez Alos (ECMWF), Matt Mayernik (NSF NCAR), Rachel McCrary (NSF NCAR), Seth McGinnis (NSF NCAR), Marion McKenzie (Colorado School of Mines), Joana Miguens (EUMETSAT), Andy Newman (NSF NCAR), Thomas Nicholas ([C]Worthy), Jonathan Petters (Virginia Tech), Amy Quarkume (Howard University), Rahul Ramachandran (NASA), Deanesh Ramsewak (The University of Trinidad and Tobago), Douglas Rao (North Carolina Institute for Climate Studies), Remata Reddy (Jackson State University), Rob Redmon (NOAA NCAI), Luz Rivera (Elizabeth City State University), La Tasha Roberts (Austin Community College), Mia Robinson (Jackson State University), Ulrike Romatschke (NSF NCAR), Evan Rusackas (Preset.io), Doug Schuster (NSF NCAR), Suman Shekhar (Rutgers University), Tasha Snow (Colorado School of Mines), Negin Sobhani (NSF NCAR), Karen Stocks (Scripps Institution of Oceanography), Jed Sendwall (Radiant Earth), Francis Tuluri (Jackson State University), Kevin Tyle (University at Albany, SUNY), Mara Ullao (NSF NCAR/Northwestern University), Jon Vandegriff (Johns Hopkins University), Kristina Vrouwenvelder (American Geophysical Union), Jacquie Witte (NSF NCAR)

Workshop Steering Committee Members: Angel Alos (ECMWF), Shanice Bailey (Columbia University), Michael Bell (Colorado State University), Scott Collis (Argonne National Laboratory), John Clyne (NSF NCAR), Teagan King (NSF NCAR), Matt Mayernik (NSF NCAR), Doug Schuster (NSF NCAR), Douglas Rao (North Carolina Institute for Climate Studies), Francis Tulari (Jackson State University), Jacquie Witte (NSF NCAR)

Appendix II - Workshop agenda

Color Key
Yellow - Breaks
Green - Shuttle Schedule
Blue - live streamed

Wednesday, May 29, 2024		
Time (Mountain Time)	**Agenda Item**	**Location (at NSF NCAR Mesa Lab unless otherwise stated)**
11:30-13:00	Lunch on your own	On your own - *the NSF NCAR Mesa Lab Cafeteria has food available for purchase from 11:30-13:30.*

12:45	Bus departs Hilton Garden Inn for NCAR Mesa Lab	Hilton Garden Inn *2701 Canyon Blvd, Boulder, CO 80302*
13:00	Workshop attendee arrival/check in	Mesa Lab Lobby
13:30	Welcome and summary of workshop structure and goals Link to recording	Main Seminar Room
13:45 - 15:15	Plenary Session 1 - Big Ideas and Existing Solutions Link to recording	Main Seminar Room
15:15 - 15:45	Break	Lobby
15:45 - 17:00	Breakout Session 1 - Big Ideas and Existing Solutions	Mesa Lab Breakout Rooms
17:00-17:30	Breakout Session 1 summaries/open discussion Link to recording	Main Seminar Room
17:30	**ADJOURN**	
18:00	Bus departs ML to Hilton Garden Inn	NSF NCAR Mesa Lab Parking Lot

Thursday, May 30, 2024		
Time (Mountain Time)	**Agenda Item**	**Location (at NSF NCAR Mesa Lab unless otherwise stated)**
7:00 - 8:00	Breakfast on your own	On your own - *the Mesa Lab Cafeteria has food available for purchase from 7:30-9:30*
8:00	Bus departs Hilton Garden Inn for NCAR ML	Hilton Garden Inn 2701 Canyon Blvd, Boulder, CO 80302
8:30 - 8:45	AM Welcome, Day 1 recap and intro to Day 2 activities	Main Seminar Room

	Link to recording	
8:45 - 10:15	Plenary Session 2 - Data Commons Analytics Research and Educational User Needs Link to recording	Main Seminar Room
10:15 - 10:45	Break	Lobby
10:45 - 12:00	Breakout Session 2 - Data Commons Analytics Research and Educational User Needs	Breakout Rooms
12:00 - 12:30	Breakout Session 2 summaries/open discussion Link to recording	Main Seminar Room
12:30 - 13:30	Lunch on your own in the NCAR ML cafeteria	Cafeteria
13:30 - 15:00 (now 13:45 - 15:15)	Plenary Session 3 - Data Producer Needs for Data Curation and Consulting Services Link to recording	Main Seminar Room
15:00 - 15:15 (Now 15:15-15:35)	Break	Lobby
15:30 - 16:30 (now 15:35 - 16:35)	Breakout Session 3 - Data Producer Needs for Data Curation and Consulting Services	Breakout Rooms
16:35 - 17:00	Breakout Session 3 summaries/open discussion Link to recording	Main Seminar Room
17:00	**ADJOURN**	
17:30	Bus departs ML to Hilton Garden Inn	NSF NCAR Mesa Lab Parking Lot

Friday, May 31, 2024		
Time (Mountain Time)	**Agenda Item**	**Location (at NSF NCAR Mesa Lab unless otherwise stated)**
7:30 - 9:00	Breakfast on your own	On your own - *the Mesa Lab Cafeteria has food available for purchase from 7:30-9:30*
8:30	Bus departs Hilton Garden Inn for NCAR ML	Hilton Garden Inn 2701 Canyon Blvd, Boulder, CO 80302
9:00 - 9:30	Breakout session 3 report out. Upvote and discuss outstanding questions that will be used to drive the final breakout session. Link to recording	Main Seminar Room
9:30 - 10:15	Breakout Session 4 - Topics based on workshop participant slido suggestions and voting	Breakout Rooms
10:15 - 10:45	Breakout session 4 summaries/open discussion Link to recording	Main Seminar Room
10:45 - 11:00	Discuss next steps Link to recording	Main Seminar Room
11:00	**ADJOURN**	
11:30	Bus departs ML to Hilton Garden Inn	NSF NCAR Mesa Lab Parking Lot

Appendix III - Plenary Sessions - Detailed Agendas, Titles and Bios

- **May 29, 13:45 - 15:15: Plenary Session 1 -Big Ideas and Existing Solutions**
  - Moderator: Doug Schuster -**Workshop Introduction Slides**
  - **5 x 15 mins**
    - **Can you see me now: Mitigating Data Pollution**
      - Amy Quarkume, Howard University

        - **Open science as a process, not a product: Insights from the CryoCloud JupyterHub communities**
            - Tasha Snow
        - **New horizons for the Data Stores Engine at ECMWF**
            - Angel Alos
        - **What might a data commons learn from the rise of the open-source/open-science software commons?**
            - Deepak Cherian, Earthmover PBC
        - **From Petabytes to Insights: Tackling Earth Science's Scaling Problem**
            - Dr. Rahul Ramachandran, NASA's Marshall Space Flight Center (MSFC)
    - 15 minutes for discussion at the end

- **May 30, 8:45 - 10:15: Plenary Session 2 -Data Commons Analytics Research and Educational User Needs**
    - Moderator: Jacquie Witte
    - **5 x 15 mins**
        - **Data data everywhere, but not a byte to process**
            - Kevin Goebbert, Valparaiso University
        - **Finding the Common Denominator: User Needs for Research and Education**
            - MaKenna Collins, Jackson State University
        - **Machine Learning Data Challenges across the Pipeline**
            - David John Gagne, NCAR
        - **The Role of APIs in Distributing and Sharing Data: the HAPI Example from Heliophysics**
            - John Vandergriff, Johns Hopkins University
        - *Data Commons: a Caribbean Perspective on User Analytics Needs for Research and Education.
            - Deanesh Ramsewak, University of Trinidad and Tobago
            - **Not livestreamed/not recorded/slides not available**
    - 15 minutes for discussion

- **May 30, 13:30 - 15:00 (Now 13:45 - 15:15): Plenary Session 3 -Data Producer Needs for data curation and consulting services (facilities and individual researchers)**
    - Moderator: Scott Collis
    - **5 x 15 mins**
        - **Data curation needs for large and small NSF facilities**
            - Asti Bhatt, SRI
        - **Data Storage and curation needs for high-resolution continental scale climate impact studies**
            - Andrew Newman, NCAR

        - **Perspectives on Data Producer Needs for the Madrigal Database**
            - Katherine Cariglia, MIT Haystack Observatory
        - **Transforming Data from Uncommon to Common: Perspectives from a Data and Software Provider**
            - Michael Bell, Colorado State University
        - **Modernizing Data Stewardship to Enable Data-Driven Environmental Research and Applications**
            - Douglas Rao, North Carolina Institute for Climate Studies
    - 15 minutes for discussion.

*Livestream and recording not available

Plenary Speaker Bios

Rahul Ramachandran

Dr. Rahul Ramachandran is a distinguished Senior Research Scientist at NASA's Marshall Space Flight Center (MSFC) and leads the Inter-Agency Implementation and Advanced Concepts (IMPACT) team. His research interests span a range of topics, including data science, informatics, and AI/ML. Dr. Ramachandran has numerous peer-reviewed publications and has made significant contributions to improve the way we manage and analyze large geospatial datasets leading to a better understanding of our planet and its complex systems. He has held editorial positions in different journals. Dr. Ramachandran is the recipient of numerous accolades and honors, including the Presidential Early Career Award for Scientists and Engineers (PECASE) and the NASA Exceptional Achievement Medal. Dr. Ramachandran was the American Geophysical Union's 2023 Greg Leptoukh Lecture recipient in recognition of his significant contributions to informatics, computational, or data sciences through research, education, and related activities.

Angel Alos

As a member of the Forecast and Services Department at ECMWF, Dr. Angel Lopez Alos is currently Leading the Common Data Stores (CDS) Team whose main task is the implementation and operational management of the Copernicus Climate and Atmosphere Data Stores. Before joining ECMWF, he was part of the Data Specifications Team for the EU INSPIRE Directive at EC-JRC. Since its early beginnings, his professional career developed around Information Technologies applied to different sectors. Dr. Lopez has a background and PhD in Earth Sciences and holds a Master in Environmental Management and Engineering by EOI.

Tasha Snow

Tasha is a Research Scientist with NASA Goddard Space Flight Center and University of Maryland working to better understand Greenland and Antarctic ocean and glacier change and how it will impact the planet. She specializes in remote sensing, open science, and cloud computing. She is a co-founder of the CryoCloud cloud computing JupyterHub (cryointhecloud.com) that aims to help usher cryosphere research communities into the virtual cloud and build community best practices for modern scientific workflows.

Deepak Cherian

Deepak Cherian, Ph.D is a Forward Engineer at Earthmover PBC working to build a data lake platform for multi-dimensional array data in the cloud. Over the past decade, he actively worked in the open-source scientific Python community to advance the limits of whats possible with the Xarray/Dask/Zarr suite of software (colloquially known as the "Pangeo stack").

Amy Quarkume

Dr. Amy Yeboah Quarkume, is an Associate Professor of Africana Studies and Data Science at Howard University. She holds multiple degrees in African American Studies, Sociology and Data Analytics. She serves presently as the Director of Graduate Studies for the Master's Program in Applied Data Science and Analytics, advancing Howard University's first major effort in becoming a hub for data science social justice research. Her work as a data scientist centers around AI Bias, data inequality, and environmental justice. Furthermore, she is the PI of the CORE futures lab, PI in the NOAA Cooperative Science Center in Atmospheric Sciences and Meteorology (NCAS-M), and NCAR Innovator Fellow.

Kevin Goebbert

Dr. Kevin Goebbert is a Professor of Meteorology at Valparaiso University where he has taught undergraduate students since 2009 with a focus on synoptic meteorology and meteorological computing. Recently he developed an online educational resource entitled “Introduction to Weather Technology using MetPy”, which is designed to develop basic Pythonic computer coding skills alongside meteorological data analysis and visualization. He has served on NSF Unidata governance committees for the past ten years contributing guidance on the development of software tools and data access and distribution for the benefit of University research and education efforts.

MaKenna Collins

Rising senior at Jackson State University, scheduled for graduation in Spring 2025, with a focus on Earth System Science and a minor in Sociology. Engaged in diverse projects from creating a Jupyter notebook for retrieving NOAA tide gauge data to studying severe weather and climate impacts in the Southeast region of the US and analyzing a rock core from the K-T boundary. Eager to apply this interdisciplinary background to other Earth science tasks and contribute to environmental success.

David John Gagne

Dr. David John Gagne II is a Machine Learning Scientist II and head of the Machine Integration and Learning for Earth Systems group at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. His research focuses on developing machine learning systems to improve the prediction and understanding of high impact weather and to enhance weather and climate models. He received his Ph.D. in meteorology from the University of Oklahoma in 2016 and completed an Advanced Study Program postdoctoral fellowship at NCAR in 2018. He has collaborated with interdisciplinary teams to develop and evaluate machine learning systems for high impact weather and emulation of atmospheric processes. He is a senior leader of the NSF AI Institute on Trustworthy AI for Weather, Climate, and Coastal Oceanography (AI2ES) and senior personnel for the NSF LEAP Science and Technology Center. He has led summer schools, short courses, and hackathons on AI for Earth System Science, chaired the American Meteorological Society Artificial Intelligence Committee, and serves as an editor for the journal AI for the Earth Systems.

John Vandergriff

Dr. Jon Vandegriff is Principal Professional Staff at the Johns Hopkins Applied Physics Lab. He supervises a group of 50 RSEs, scientists and engineers who create tools and perform analysis on space mission data covering mainly Heliophysics and Planetary Science. He coordinates the development of the Heliophysics Application Programmer's Interface, a grass-roots community effort standardizing access to time series data.

Deanesh Ramsewak

Mr. Deanesh Ramsewak is an Assistant Professor in Practice and leader of the Bachelor of Science Programme at the University of Trinidad and Tobago's Centre for Maritime and Ocean Studies. He teaches GIS and Remote Sensing and his research focuses primarily on the use of machine learning techniques for mapping and monitoring of coastal ecosystems such as mangroves and coral reefs. He has worked on many Caribbean Sea initiatives supported by NASA, the European Space Agency and the International Oceanographic Commission, and was recently appointed as a member of the UN Ocean Decade - Digital Twins of the Ocean Steering Committee. Mr Ramsewak is also a Fellow of the Royal Geographic Society and the Institute of Marine Engineering Science and Technology, and is a Science Mentor for the Space Generation Advisory Council (SGAC) in Vienna, Austria.

Michael Bell

Professor Michael Bell has expertise in tropical weather and climate, field observations, and remote sensing. He is the principal investigator for the CSU SEA-POL Sea-Going and Land-Deployable Polarimetric radar, which is an NSF Community Facility. He is also one of the leads of the Lidar Radar Open Software Environment (LROSE) project in collaboration with NCAR EOL.

Katherine Cariglia

My name is Katherine Cariglia and I am an undergraduate student studying computer science at UMass Lowell. I started working at the MIT Haystack Observatory when I was in high school, and I focus mainly on maintaining the Madrigal Database.

Andrew Newman

Andrew 'Andy' Newman has Atmospheric Science B.S. and M.S. degrees from the University of North Dakota and his Atmospheric Science Ph.D. from Colorado State University. He has been at NCAR since 2011 where he focuses on developing actionable Earth science for partners spanning water, climate, and human health across spatiotemporal scales.

Asti Bhatt

Dr. Asti Bhatt is a research scientist from SRI with over 10 years of experience in the field of ionosphere-thermosphere-magnetosphere science. She is the PI of two NSF facilities – the MANGO network of all sky imagers and the AMISR radars, which produce near-continuous data for the broad geospace community use. She is interested in ensuring good data stewardship of geospace data, which includes computational reproducibility and efficient use of vast and varied data through modern data science tools.

Douglas Rao

Douglas Rao is a research scientist with North Carolina Institute for Climate Studies and affiliated with NOAA National Centers for Environmental Information. His research focuses on leveraging innovative technologies to enhance the value of climate data for impact studies for ecosystems and environmental health. He currently serves as the Vice President for Earth Science Information Partners and leads a cluster in developing community standards for AI-ready open data.

Appendix IV - Breakout session discussion questions by workshop theme

Wed, May 29, 15:45 - 17:00 Breakout session 1 -Big Ideas and Existing Solutions

- What is your domain expertise?
- Reflections on the plenary session talks?
- Describe a “data driven” research or educational use case (e.g. compute statistics or derived datasets, produce plots, train AI/ML models).
    - If you currently use a “data analytics platform” (e.g Science Gateway with JupyterHub) to support your research or educational use case, what is it and what capabilities are provided to enable your research or educational use case?
- Are there any gaps that exist in either of these environments that inhibit progress on your research or educational use case?
    - What are some challenges for you to access these environments?

Thu, May 30, 10:45 - 12:00 Breakout Session 2 -Data Commons Analytics Research and Educational User Needs

- What is your domain expertise?
- Reflections on the plenary session talks?
- Do you have any data driven research or educational use cases (e.g. compute statistics or derived datasets, produce plots, train AI/ML models) that you currently can’t do or are inhibited in accomplishing due to resource constraints?
    - What are the specific barriers that inhibit your research or educational workflow for your use case?
    - What capabilities would make it possible for you to explore your research or educational workflow in a more effective manner?

Thu, May 30, 15:15 - 16:30 Breakout session 3 -Data Producer Needs for Data Curation and Consulting Services (facilities and individual researchers)

- What is your domain expertise?
- Reflections on the plenary session talks?
- What type of datasets do you produce and do you produce them as an individual PI or through a community facility?

- Are you expected to make your datasets publicly available to support funder or publisher requirements? If yes, how do you currently do so?
- What challenges do you experience when attempting to publicly share your dataset?
    - What capabilities would help you overcome these challenges (Consulting and technical capabilities)?