Thematic Challenges and Recommendations

Throughout the 2023 Workshop on Open Data and Reuse in Social Science Weather Research, participants were invited to share challenges and concerns related to the Office of Science and Technology Policy (OSTP) guidance, as well as recommendations for overcoming these challenges. This document offers a thematic summary of what was shared in writing and verbally during the meeting.

Managing and Curating Data Instruments

Challenge 1: Leveraging Data Repositories

- How can NOAA (or other agencies) leverage work done with data repositories, such as DesignSafe, Dataverse, ICPSR, QDR, etc.? How can we collaborate more?
    - How do the landscapes of repositories need to change?
    - What does this research community need?

Recommendations for Challenge 1

- Encourage our PIs to identify their repository early and include them in the transition plan discussion (with respect to data as an output). This allows the PI, Repository, and NOAA to discuss and respect data sensitivities.
- Repositories could include the pending DOIs for the grant so we can track data for grants.
- Feds would like to learn more about each repository and understand how their search functions work and learn more about their metadata process.
- Co-create together and create a set of keywords that a few repositories may use uniformly? (NOAA, weather, disaster, etc.) *Consider creating or using a published controlled vocabulary to standardize some searchable metadata fields.*

Outstanding Considerations for Challenge 1

- Data management policies (e.g., requirements, timeline) differ across agencies. What can be done about not only having more common policies but also policies that make it easier to publish data and instruments?

Challenge 2: Definition of "Data" and Data Formatting

- Would coding in any software, such as Python, R, GitHub, etc. be considered a data product? More broadly, how are we defining data in this space?
    - Should there be an expectation that code is included in the dataset for it to be FAIR or sufficiently integrated, usable, and understandable.

Recommendations for Challenge 2

- Clear definition of terms, what is considered "data"/ "data product" (Including code in those definitions)
- Research software can be considered its own publication or can/should be added to a data publication as analysis or documentation files.

- Data repos need to develop standards and guidance for defining what programs can be elevated to "software publication".
- DesignSafe is developing that policy and has a definition for software publication and a metadata profile already for software pub.

Outstanding Considerations for Challenge 2

- What can we do to develop a standard of formatting for data files, such as systematic headers, data contextualization, etc.?
- What level of granularity do you index data? Does this matter for different data types?
- Is there a common understanding, or common standards, of what constitutes "curated data"?
- What is or should be the goal of curation? What is the minimum level of curation necessary to support FAIR principles?
- What does open social science data look like? Transcripts? CSV files? Etc.
- What if, considering the OSTP guidance and the rush to publish, data publications might be scattered across several data repositories that might be related. What are the implications of that? Do we want to land on one repository? Is there a strength to spreading out?
- How does one decide on what type of repository to share their data with? There are differences in domain-specific, generalist, or institutional repositories. Should there be federal standards on how to decide?
- ICPSR, Dataverse, and DesignSafe represent three different models of data repositories, each with slightly different practices. Is there, should there, be one that rules them all, or are there paths forward where all three co-exist with shared goals but supporting different communities (but the repositories can communicate and share with one another)? Is GREI a model for this?
- Which data format(s), metadata, and weighting schemes do we need to invest in sharing? Different sets of weights can change the data, do both and all these need to be shared? What counts as data? Do weighting count, as they tend to represent small populations?
- How should we (digital and analog) infrastructure supporting research workflows feed into infrastructure primarily designed for research data preservation? How do we publish and analyze "event bundles" even when it is in a workflow infrastructure and other types of very structured data? What if it isn't digital? How should they interface with repositories and those designed for data research preservation?
- (How) should qualitative data collected in the field be shared? What about the (meta)data that is in our minds and bodies? How can we document things collected in the field? How do we capture the tone of voice, the feelings and body language from people, and the contextual information in a meaningful way that doesn't identify people? How to handle the relationship between researcher and informant and the trust and relationships implicit in the data collection?

Challenge 3: Effective Documentation and Metadata

- How do we create effective documentation (e.g., what's in it and how it gets to others), including metadata, lab notes (*terminology for lab notes: field notes, work logs, lab meeting notes) and others, that can facilitate use broadly across researchers and organizations and ensure replicability?
    - What kind of metadata is appropriate?

Recommendations for Challenge 3

- Affordable infrastructure for "living" data.
- Encourage starting early.

- Create templates or standards/analytic frameworks—e.g., for field notes, lab notes / work logs /; Prompts—don't forget to (tell people your data types); recognize limitations and provide affordable infrastructure for hosting documentation.
- Version control - support sharing of techniques across fields for versioning (including versioning in iterative study design). Attention to directory design, folder structure, and file-naming schemes (international (ISO) standards) Documentation bundled with datasets. Designated responsible party for oversight of documentation.
- Consider documentation as a possible approach to reducing bias *or* creating interpretive intensity. Metadata on intentions.
- Caution: Pay attention to power structures in the work and how that affects lab notes etc. (e.g., consideration of who to include (named students) in lab notes in case of abuse from faculty or PI and potential FERPA violations. Resource: Turing institute guidelines for reproducible research and creating collaborative work).

Outstanding Considerations for Challenge 3

- How can we develop a standard of formatting for documentation of metadata?

Publishing Data and Instruments

Challenge 4: Working with Teams/Collaborators

- Collaboration is wonderful, but it brings many challenges. Surveys or data can be joint intellectual property and to go back and share that data, how do you get to decide who shares? Who holds that responsibility? If people get busy, will they get back to doing those tasks? There are situations where it is clear, but is that person going to follow all the way through with data publication? At other times it might be very unclear, and then it becomes more complex.
    - How do we design principles and agreements for co-authorship or credit upon releasing data for public use, and ensure compliance with your agreements?
    - How do you negotiate sharing and publication when the data/work is the fruit of collaboration?
    - **How do you decide who is responsible for sharing data for reuse, when you are part of a collaborative project?**
    - **How do you negotiate sharing and publication when the data/work is the fruit of collaboration?**
    - **How do we ensure that researchers who develop and produce original data relative to publications are credited and how are user agreements incorporated into this process?**
    - Should data creators be listed as co-authors on a paper that reused their data? Is citation insufficient?
    - **What does this (social science) research community need?**

Recommendations for Challenge 4

- Emphasize data publication in funding opportunities and proposals and address lack of clarity. This requires training reviewers to value (and demystify) data management and publication.
    - Perhaps the location of the data management (and sharing and reuse) section in proposal structure causes people to see it as an add-on (and unfunded mandate) at the end, whereas it should be thought of at the outset of project design with collaborators.
- Shared solution: early project design with archivists and librarians to co-create data plans.

- Partnership agreement created at the outset of the study/proposal for how data will be shared and published. Iterative conversations with collaborators going over this template/list of best practices. The goal will be an agreed upon strategy for sharing and publishing data.
- Co-authorship creates a mechanism for communication and correspondence mechanism for asking questions about data collection and curation and context.
- Should data creators be invited to collaborate? Sounds great in theory, but if a data resource becomes popular, this is not a sustainable policy.
- Is utilization and citation insufficient?
    - Co-authorship gets more credit. How do we elevate data publications to true first-class research outputs? Does it depend on reuse? Citation counts? Do we have members of the hazards and disaster social science field who are big "data reuse" researchers to help guide in this endeavor?

Challenge 5: Working with Communities

- What role can community leaders or small community-led organizations play in decisions regarding their data? There is data we could publish, but with community partners it could hinder their lives. What if another researcher is interested but they decide to publish it? What do long term concerns play? What happens if they don't want to share their data? How do we navigate this process with them? How do we de-identify their information, but considering the size of the community, it will be easy to find out who they are. How can we build trust?
    - When do we need to return to the community? How do we get this information to them under this guidance? How do we use our data to give back?
    - How can we create guidelines for data reuse by communities to encourage data reuse for advocacy purposes?
    - What role can community leaders or small community-led organizations play in decisions regarding their data reuse?

Recommendations for Challenge 5

- Do we need to have community workshops and meetings? Indicator workshops with community members could be useful in understanding their own concerns around the data collection and publication process.
- Need to consider and be cognitive of power dynamics.
- Potentially have robust templates for partnership development. This facilitates the process for researchers who want to do this kind of work, and they may not know where to start.
- Three-way data use agreement (DUA) between communities, researchers, and prospective users of the data.
- Have an NSF Research Coordination Networks (NSF-RCN) or Research Traineeship Program (NSF-NRT) as well for addressing the issues upstream.
- Staff for support.
- In the data sharing, write how the data can NOT be used for a community agreement. This is one way of overcoming the challenge of unknown ways in which it may be used.
- There should be more guidance on what is made available and how it works—the process. Not just substantive data points. The Federal mandate should focus not just on data points as the output but *the processes* that are published and released to work with researchers to use that data ethically.
- Time and returning this information to the community.

- A community partnership agreement, or something similar, could also address this issue if we agreed upon outputs—what would be useful to the community? For their decision-making, planning, etc.

Challenge 6: Impacts on Security and Sensitive Data

- There are concerns about how the new OSTP guidance impacts the collection of security and identity aspects of data!
    - 1. Personally Identifiable Information (PII), Personal Health Information (PHI), Health Insurance Portability and Accountability Act (HIPAA); and data collected from communities that may be difficult to de-identify or who may be reluctant (or unable) to share because of later downstream consequence(s).
        - How do we continue to overcome reluctance to share data? Are there specific interventions needed?
    - 2. Controlled Unclassified Information (CUI), Protected Critical Infrastructure Information (PCII), other types of difficult to de-identify data or data that simply can't be collected unless certain management/sharing protocols are followed.
        - Should users have special credentials to reuse sensitive data?
    - How do we comply with federal mandates while also protecting these populations?
        - Privacy concerns with sharing data
        - What is the tradeoff between public access and privacy?
        - Personally identifiable (PII, PHI, HIPAA)
        - Data with security issues (CUI, PCII)

Recommendations for Challenge 6

- Office of Human Research Protections (OHRP) buy-in and procedures/policy for IRBs and researchers (to give permission to academic researcher, address compliance concerns of institutions, and show the clear path/process)
- Separate data collection (instrument) from data reporting
- Access to data is clearance dependent. Requests for access are evaluated based on the type of "public" making the request (academic research, state agency, community-based organization, individual).
    - For example, Data.gov repository for government agency data evaluates "public access" applications for access to data to ensure security. (Similar to how you apply for access to a limited use dataset)
- Some repositories including ICPSR, Harvard Dataverse, QDR and DesignSafe make a distinction between publishing and secure access to authorized users. Repositories offer different levels of access depending on how sensitive and disclosive the data are. There is an agreement—which takes different forms depending on the repository, from signed agreements to interviews—between the depositor (individual and institution) and the repository, that specifies the level of restriction for that data set. For some repositories the access restriction comes to the file level, for others it comes to the dataset level. Some repositories handle restrictions upon the agreement with the institution or individual, others allow depositor-approved access. Identifying the type of restriction for a particular collection depends on the IRB-approved informed consent, the analysis of the data products in relation to a disclosure risk review, the policies of the repository in relation to publishing diverse forms of identification (direct, indirect).
- Repositories also provide curatorial guidance to publish aggregated forms of the data or redacted data.
- Differential privacy tools are being developed to release sensitive data with added noise that reduces risk of re-identification: https://opendp.org/about

Outstanding Considerations for Challenge 6

- If sensitive data is shared via repository to NOAA, who would be the authorized representative within NOAA? How would that change as people, roles, and capacities change?
- What is the impact of the Freedom of Information Act (FOIA) on data sharing of sensitive data?
- How do we manage [conflicting] sensitive data demands in federal funded projects? How do we address non-disclosure agreements, their vulnerabilities, and who gets to see these? Who gets access? How does that get managed? How do we make sure agencies and others are comfortable with the process that (should) ultimately benefits them? How do we account for that in the budgeting process?
- What if IRB guidance conflicts with funding requirements? How do we get the complicated process of working with other institutions in the data publication requirements? It can be one more thing to add to the consent form, and then those who are being interviewed do not know what this means. How are we writing our protocols? How do you work with funding agencies for these requirements? How do we build university standards?
- How do we handle a "mix" of data, published or not? Federally funded or not? Who holds the responsibility and rights?

Challenge 7: Guidance and Training

- If funders begin requiring more extensive data management, sharing, and reuse plans, what templates, guidance, or training will researchers and data archives need?
    - What are the tools available (workspace etc.) and training for data reuse?

Recommendations for Challenge 7

- Funders and federal agencies will need to understand the current OSTP guidance, laws, acts etc. (top down) but they will also need to consult with and listen to communities to understand needs, concerns, etc. from the researchers, publishers (journal editors), and the data archives (bottom up) to make sure relevant information is captured.
- Federal agencies, researchers, publishers (journal editors), and data archives will need to co-develop trainings, templates, or other materials to help researchers to do this well and with an *equity* lens (and remember that anything created—if adopted at the federal agency level— will need to be reviewed and approved by federal agencies.) *Could be useful to work with institutions and libraries to add content to training that features local/institutional resources for researchers.*
- Socialization - there will need to be time for members of the social science research and interdisciplinary research community to learn about the new requirements and to contribute to the templates and training.
- Review Process - When the federal agencies and panels begin review processes, there will need to be additional guidance at the outset of the panels that these "new" data management, sharing, and reuse plans are submitted, they are taken seriously, and that applicants for awards will receive feedback on the quality and focus on equity of submitted plans. Also, the legal language in calls for proposals will dictate the level of legal review needed for any resources provided/required. *Can agencies develop rubrics (in addition to templates) for evaluating Data Management and Sharing Plans that can guide researchers on writing a high-quality plan?*
- Timeline: There may need to be a *strongly encouraged* period between now and FY27 when the full changeover may take place over may likely *require* these plans. See **Socialization** section.
- Dmptool (or other tools) are available for the benefit and use of the community.
- NSF already has good descriptions for the data management plan (and it can only be two pages). Researchers may need to be socialized to talk more about equity, data sharing, and reuse, etc.

Outstanding Considerations for Challenge 7

- NOAA does not currently have extensive guidance.
- How do we take this conversation into the classroom and into mentoring spaces? How do we help mentor graduate students to become the next generation of data managers and publishers BEFORE they move to federal jobs or academic positions?
    - training students/research groups for future projects
    - using the data in class to teach students.
- What is/should be the role of academic libraries in data sharing and publication? Education, assistance, repository?
- How do organizations guide users about data reuse?
- Data sharing and publication relies on data curation skills and services, but this is still a very nascent field. Who is the focus of education for this field? What is the goal?

Sharing Data and Publications

Challenge 8: Finding Existing Data and "Legacy" Data

- How do we figure out where existing data lives, how can we access it, and make it more readily available through searches or other "data finding" projects?
    - Lookers = Primarily researchers? Data librarians? With some interest from both funders and repositories.

Recommendations for Challenge 8

- Repositories should create Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) collections & sets for other catalogs and repositories to harvest.
- Could library catalogs have a role in collecting data records from diverse repositories?
- DataCite catalog

Outstanding Considerations for Challenge 8

- How do we deal with publications relative to the relinquishing of the existing data embargo? How will this affect legacy work versus new research?
- How do we get legacy data to reuse?
- What are the requirements to permit datasets alignment and reuse?
- Funding agencies may lose purview to oversee data sharing past a certain (temporal) point. How can we encourage sharing within the appropriate time frame while being responsible for the protocol of the OSTP guidance?

Challenge 9: Considering Publication and Publishers

- Publications are the usual end goal of data collection and analysis. How do we take the considerations of publication into account when it comes to data publication, management, and sharing?

Recommendations for Challenge 9

- Encourage publication of articles about data—for now, this might be more of an incentive for academics that don't (currently) get much credit for publishing data alone.
- Formally cite data that you use/consult in your publications.

Outstanding Considerations for Challenge 9

- Publications might fall under the purview of open data within OSTP. This is not the current culture of publishing, as open access is currently costly. What is the role of the publishing industry moving forward? Should they have a role at all?
- How can we leverage persistent identifiers like DOIs, ORCID, and others in data publication and reuse?

Incentivizing Data Creation, Publication, and Reuse

Challenge 10: Incentivizing Data Reuse

- How can data reuse be cultivated and encouraged? How do we incentivize data reuse?
    - Research funding and federal agencies perpetuate a culture of primary data collection, instead of secondary data sharing/use. How do we incentivize this work?
    - Who is the audience that will reuse the data?
    - What are the funding opportunities to reanalyze data?
    - What is it that on the federal side does to encourage people to use and reuse data? What do we need to do on the federal funding agency side?
    - How will federal agencies act as promoters of data sharing and start this process?

Recommendations for Challenge 10

- Recognize the need for cultural change, and both institutional and technical support. This could include:
    - Actively share data available for re-use, maybe with weekly or monthly newsletters consisting of a title and very brief (clickable) description. Also share examples of interdisciplinary "integrated" data projects. Design-Safe already hosts data re-use stories.
    - Create technically supported workflows for commenting on datasets so that re-users feel supported by other analysts in their assessment of the data set.
    - Make clear (through statements from funders, for example) that data re-users are not themselves culpable (intellectually or morally) if a dataset they re-use is later found to be less reliable than they thought when they decided to use it.
    - Recognize that communities of practice need to be built around data re-use. A step in this direction are the research groups and thematic/topical collections some of the repositories have built.
    - Have user protocols or 'what not do to' files.
    - Readme files clarifying a community's willingness to help (resource permitting) (And WATCH ME files!) (This is a chance, potentially for data authors to clarify how they do or do NOT want to be contacted, involved, etc.) moving forward.
    - Don't want to restrict data use, but also don't want to be involved in the whole thing and occupy much of your time.
    - Re-use should be accompanied with resources. (There could be user forums like ICPSR does for very popular datasets that are often reused)

- Research award programs (funding!) or Paper competitions for undergraduate/graduate students.

Outstanding Considerations for Challenge 10

- The integration of data of different formats, time scales, privacy concerns, etc. for effective reuse…how do we reuse heterogeneous datasets?
- What are the minimum standards and attributes of databases to permit data reuse?

Challenge 11: Incentivizing Interdisciplinary Data Collection and Reuse

- How could you make or create new interdisciplinary datasets?
  - How might government agencies incentivize or motivate the creation of interdisciplinary datasets using published physical and social science data?
  - What are the research questions that would produce important new knowledge and how can you use existing datasets to answer those questions?
  - How do we reuse interdisciplinary data? How might government agencies incentivize or motivate researchers to reuse or create interdisciplinary datasets?

Recommendations for Challenge 11

- Physical and social science data are not always located inside the same repository. Create linkages needed to create new interdisciplinary datasets.
- Create metadata and data standards that enable integration on dimensions of importance for key questions across sciences (e.g., temporality, spatial dimensionality, social, connections/networks, intersectional characteristics).
- Peer review of data reuse may be needed to guard against data misuse, especially for interdisciplinary data.
- Interdisciplinary users need more information in the repositories on the value of these data to articulate how this dataset may help answer a novel question. The creation of one pager, clearer descriptions, and plain-language data value statements.
- Interdisciplinary repositories might benefit from features that allow users to create profiles that allow/enable them to collect data vs. simply using a profile to upload data.

Outstanding Considerations for Challenge 11

- We cannot (may not be able to) just download interdisciplinary data because it's too big, it needs to be computed in the cloud. Different access points, questions of differential access by sector for example. Usage fees for cloud storage/cloud computing—lots of questions to address about this.
- How can we combine/integrate data to make interdisciplinary datasets.
- Reuse: PI perspective—if you publish the data, it is primary for you, for others it is secondary; reuse of federal datasets is standard for many kinds of social science data—e.g., general social survey. Must be explicit about this perspective.
- What are the research questions that produce important new knowledge and what data from what sources are most informative to address those questions?
- Qualitative data are hard to put in these spaces for multiple reasons; governance data can disappear as circumstances change (decision making)—hard to reuse and shouldn't be. Often kinds of data are quantitative or analyzed qualitatively. Often social science data become "one-noted"—treated as a single thing when there are many kinds of social science data. How can we develop a way to be inclusive of all data types?
- For interdisciplinary data, need metadata and data standards that enable integration on dimensions of importance for key questions across sciences (e.g., temporality, spatial dimensionality, social,

connections/networks, intersectional characteristics; question order etc.—structure of data is important in different ways for different sciences).
- Hard to access interdisciplinary data that are not housed in the same repository. How do we link across repositories?
- How do you scaffold or build a system where you don't throw out questions that are important before you have enough data to address the questions you are interested in?
- If you are data rich, what would your vision be for people that want to request access to secondary data to answer important questions?
- How do I know if this dataset is worthy of downloading? Interdisciplinary users may need more information on the value of these data. Such as a one pager, etc.
- When it comes to an interdisciplinary repository, it would be nice to have a profile in the repository that would allow you to select and capture the data you are interested in using in analysis vs. just having a profile to publish data.
- Would some sort of interdisciplinary repository better support collaboration—as this is an important element of interdisciplinary data and research?

Data Reuse : Definitions and Practice

Challenge 12: Defining "Data Reuse"

- What is the definition of "data reuse?"
    - **What types of data can be reused? How?**

Recommendations for Challenge 12

- There is a golden standard that is talked about: reuse of the data by someone other than the original author AND for a different purpose than the original use. (Christine Borgman), but the reality is that this varies by discipline and/or sector.
- Different ways to reuse data: (1) using data as a reference; (2) data reused by its authors (not the original use, but for ongoing work); (3) reused by author AND new people for a different purpose; (4) parts of a dataset used and integrated with a new dataset; (5) an entire dataset reused for a new purpose; (6) training students/research groups for future projects; (7) using the data in class to teach students; (8) authors refreshing themselves on their own past work; (9) adding onto an existing dataset.
- What counts more? What can I (researchers) get out of it? Citing yourself.
- Data Repository: if it's referenced, by either the original author or someone else.
- It's important to acknowledge that there are different types of reuses. And the most "pristine" is not the most common. There are also expectations of reuse that need to be reconfigured. But sometimes, data is not reused—the paper is cited.
- Journals are dictating the value for much of this discussion of reuse and how it counts—along with tenure processes and NOAA/repository metrics & reporting.
- Reuse is not the same as citations. Does data have to be in a repository and downloaded to count as reuse? What everyone values for reuse differs–using it in the classroom counts just as much as for other projects for some, but not others. The value seems to be dictated by the entity. For example, academia's imposed value is for citations. PIs are incentivized by this process because the data gets better after someone else uses it.
- Impact is hard to measure, without counting things like views or downloads. That can gauge the impact of the process of sharing but doesn't dictate the value of the data or the reuse of it. But this

system of reuse counts is imperfect. Data librarians are generally skeptical of reuse; reuse doesn't talk about the quality of the data.
- Sometimes, the popularity of the data is related to the popularity of the paper.
- The definition of what "counts" can vary by discipline and lens.

Challenge 13: Data Misinterpretation and Misuse

- What do we do if people misinterpret or misuse our data?
    - Potential data misuse. What does metadata look like in this context?
    - What do we do if people misinterpret or misuse our data? How do we encourage "responsible" use of our data? How can we prevent misuse? Misinterpret? Especially when the data is used so widely. What if it is unintentional? What if it is intentional? How do we help people meaningfully reuse your data without becoming part of a bunch of projects?
    - Are there other data concerns that should be considered. Related to FAIR, should CARE principles be emphasized as well? And TRUST principles?
        - TRUST principles: https://www.nature.com/articles/s41597-020-0486-7
        - CARE principles: https://www.gida-global.org/care
        - https://www.nature.com/articles/s41597-021-00892-0

Recommendations for Challenge 13

- Community ethical codes or Legal repercussions for reidentifying and re-publishing sensitive data, PII, PHI, etc.
- How do we prevent this? Co-authorship idea—collaboration would go a long way to ensuring that data are reused appropriately. Limitations are understood by the reuse team, etc. Again, how sustainable is this?
- Is this an example of research misconduct that can be addressed via existing frameworks? (Disciplinary ethical policies, IRBs, University systems, etc.???)
- Make creating metadata and documentation easy for researchers to complete.
- Research and discipline communities need to come up with standards. Datasets for Datasheets: https://opendp.org/about and Data Nutrition Labels: https://datanutrition.org/.

Challenge 14: Unintended Consequences of Reuse

- What are the unintended consequences of data reuse and for science itself? …and for science?
    - AI is an issue; bots could be doing unimaginable things with data.
    - Data could be used wrongly (misinterpreted).
    - Data could be used for nefarious purposes (hostile powers).
    - Data collection or data publication could be limited by researchers if they think their data will be misused. While the OSTP guidance is meant to advance science, could it be unintentionally causing some research to not be done because they don't want to publish their data and have it shared.
    - How do we effectively share data such that someone else can meaningfully and ethically (re)use it?
    - Open access is not always appropriate, depending on the project, data, or other information/dimensions. If the requirement is in place, it might stop researchers due to the requirement to publish. How do we promote "as open as possible but as closed as necessary"?

Recommendations for Challenge 14

- We need to talk through data exemptions, first. Do we go for a very conservative threshold in terms of what data can and should be shared, and ultimately reused?
- There could be a layered approach to release with a project (following the workflow) where first teams create a project, then they release instruments, then they release data.
- Standards could be placed on levels of access.
- Authorship agreements are very important for data reuse to be clear if someone uses someone's data. For example, if you use someone's data, and they help with analysis, should that raise to the level of being an author? Is collecting and sharing the data alone enough to be the author. We had a very interesting conversation about the fact that it may not make sense to have an authorship agreement with the original data publishers, others may not. At bare minimum the original data must be cited. But then some people may say that they do or do not want to join as a co-author on the data, others who may say that they want to always be an author on the resultant publications…. Can the data archives offer a suite of possibilities for publication possibilities, but then there needs to be an equity lens?
- There could be download requirements.
- Limit data publication to certain repositories that have high levels of security and access.
- Could the federal agencies set guidelines or principles. The federal agencies will set a very (conservative, careful, high bar) for what data needs to be published.
- Could an unintended consequence of this policy be that this systematically disadvantages qualitative research in the review stage, their projects may be reviewed lower as they are less likely to share their data project (or be slower about it), their data may be less likely to be used because the researcher is the instrument, and each qualitative dataset is so contextualized.

Data Maintenance and Responsibility

Challenge 15: Responsibility and Maintenance of Data

- What ongoing responsibility do researchers or repositories have for the ongoing care and maintenance for files, data, etc. to facilitate/enable data reuse?
    - When all this data is published, is there going to be an expectation to help people analyze and work with this data?
    - How do we ethically share data that is publicly available at the time of the research (like twitter data), but where human subjects of the original platform later want to change availability? What if those people on social media do not want to share their data for research purposes? Can you just share how you collected the data? How do we do this with changes in social media platforms? How do we handle the data not being available after? What if it is someone else's data?
    - If unpublished data are used in a meta-analysis or review, should they be published? By whom? How can we get a coherent answer, even if they are in the process of becoming publicly available?
    - How can researchers be responsible for abiding by data sharing requirements? What are the implications for not doing so?

    - How do we ethically reuse data that is publicly available at the time of the research (e.g., Twitter), but where human subjects or the original platform later wants to change availability?

Recommendations for Challenge 15

- Ongoing responsibility for maintenance/care: Existing feedback mechanism (DesignSafe, for example) for comments or questions about datasets especially in cases where the PI cannot be reached - ticketing system for tracking until issue is resolved. Can this be the standard across the board?
- Could send reminders through DesignSafe and other repositories for annual or routine maintenance (automatic notifications) for researchers to go through data, see if it needs to be updated.
- Facilitating reuse:
    - Efforts to contextualize data within historical contexts as part of the description of the data.
    - Have some kind of written guidance for sharing your data, including describing your data. We could have a required 'context' component to publish data. So, it could be built into the infrastructure as a requirement.
    - This context could be written or recorded. "Read me or watch me" buttons.
    - Invite historians! And science communicators! Making data as accessible as possible - equity considerations.

Other Outstanding Questions

Challenge 16: Time to Publication of Data

- COVID-19 is exemplary of a cautionary tale. There are real inequities in how we interpret "immediate." But what is the support for this? Why are we prioritizing immediate over timeliness or equity? What can we learn from the pandemic?

Challenge 17: Tracking Reuse

- Should we collectively develop best practices for tracking metrics for sharing, reuse? Should there be other indicators used (not just download counts)? Should we track differences between types of downloads and other data uses?
    - How do we track data reuse? Is there a system that we can develop for this type of work?

Challenge 18: Role of Academia and Institutional Inequality

- What other kinds of things can we do in academia to help begin to normalize this culture shift? Such as valuing in the tenure process, requiring as part of dissertation, requiring as necessary for degree, or other considerations?
    - What changes need to be made to researchers' IRB practices to better prepare for data sharing/publication on the front end? Do bad practices of data reuse have consequences for the Institutional Review Board (IRB)?

    - Changes need to be made to certain academic processes (e.g., tenure, publishing) for the changes that repositories are making to be more meaningful. How do we encourage this process? What support can we provide to researchers, even as institutions that might not have resources?
- How will institutional inequities be addressed and alleviated?
    - How will the OSTP guidance be implemented in resource constrained environments and in diverse contexts? How can we help those who might not have the resources, promoting underserved research and historically underrepresented researchers?
    - Will there be additional funding to help researchers with this guidance? There are already inequities with the funding, how do we enforce equity? We have been publishing other materials as academics to help others learn and build on that and make that guidance moving forward? Peer reviewed processes for publishing data?

Challenge 19: Other Considerations

- What could be the role of AI in connecting datasets across repositories? Should there even be a role for AI in this work?
- How can data be reused to support public policies formulation and implementation?
- How do we prepare reviewers to analyze papers with data reused?
- Is there a peer review of data? Are there standards for data publishing and who is/should be the arbiter of these standards?