The Elephant in the Room

Challenges to Publishing and Reusing Social Science Instruments and Data

Session 4
Tuesday, April 11, 2023
2:45-4:30 p.m.

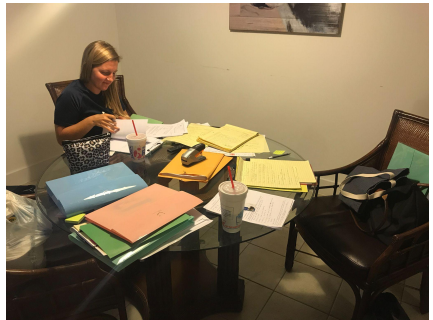
Part I: Data Curation and Publication

How do you document lab notes from your team as part of end-to-end data management to ensure replicability?

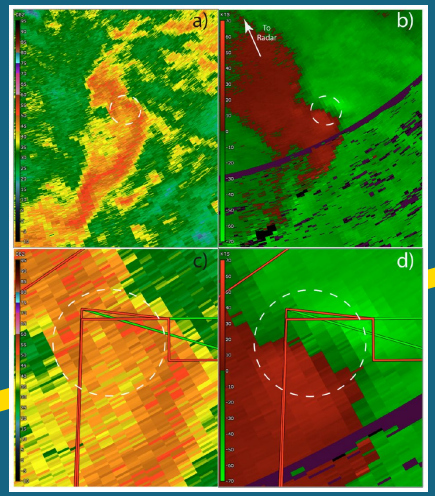
Lauren Clay

- ➔ Field work documentation (planning notes, decisions, etc…)
- ➔ Notes on the process for data management, analytic decision making, analysis (the details that go into creating data documentation, catch mistakes, incomplete work)
- ➔ Lab team meeting minutes
- ➔ RA project progress documentation (projects last years, sometimes longer that academic programs, RA contracts)

IDEA: Electronic notebook tool (e.g., LabArchives)



How do you share difficult to de-identify, small data like ethnographies?



Jen Henderson

In an interdisciplinary NOAA grant examining compound tornado and flash flood hazards in the Southeastern U.S., my meteorology colleagues and I combined analysis of the physical mechanisms of these hazards and management of warnings via ethnographic interviews with a National Weather Service Forecast Office. The specific location and type of hazard makes NWS office jurisdictions more easily discoverable. Blame and deep feelings of responsibility heighten concerns for participants about privacy.

We de-identified individuals and the office, changed dates/times/radar images so as not to reveal the location (see image), and made up a town name (Telmin). One colleague in meteorology even de-identified this case in his dissertation for additional protection. Still reviewers quickly figured out our office--we assumed readers would, too.

An N of 10-12 participants who have specific office roles means some individuals can be identified once the office / location is known. We got permission from this office to publish after they read the manuscript and gave feedback.

We face similar issues studying emergency managers and broadcast meteorologists. And with ethnographic methods.

(How) Should qualitative data collected in the field be shared? What about the (meta)data that are in our minds and bodies?

Julie Demuth

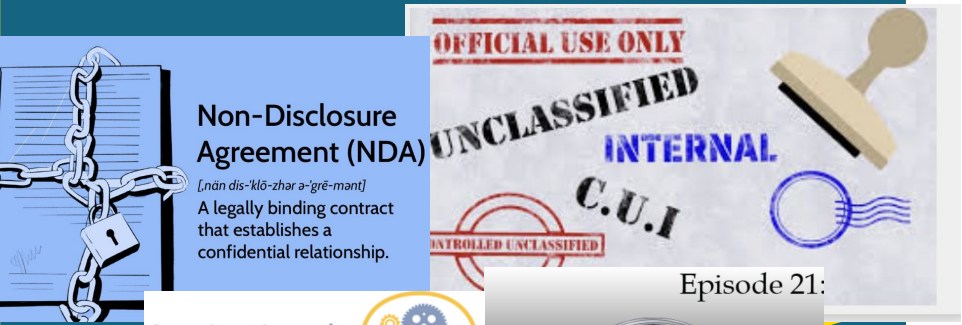
A few weeks after a deadly tornado outbreak, Heather Lazrus and I went to the areas affected to do in-person interviews with members of the public (and Jen Henderson interviewed forecasters).

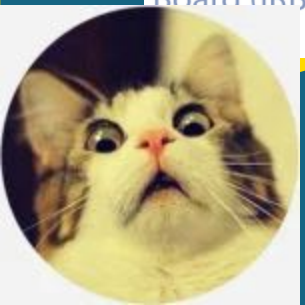
We got our interviews transcribed for analysis.

But when I analyzed the data (with tremendous help from Jamie Vickery – thank you!), I didn't just see the words that were externalized digitally. I could hear in my mind the cracks in people's voices when they shared their stories, the ways they talked faster when relaying the approaching tornado, etc. I could picture the destruction to their homes, the distance from where they lived to where they'd have to go to shelter, etc. I could feel their emotion. These (meta)data informed our analysis.

How do we effectively share those data such that someone else can meaningfully and ethically (re)use it? Is it even possible?

How do we manage [conflicting] sensitive data demands in federally funded projects?





Austin Becker (Univ. of Rhode Island)

Our Dept. of Homeland Security funded project collects data from Critical infrastructure Facilities and Military Installations for use in emergency management and resilience planning.

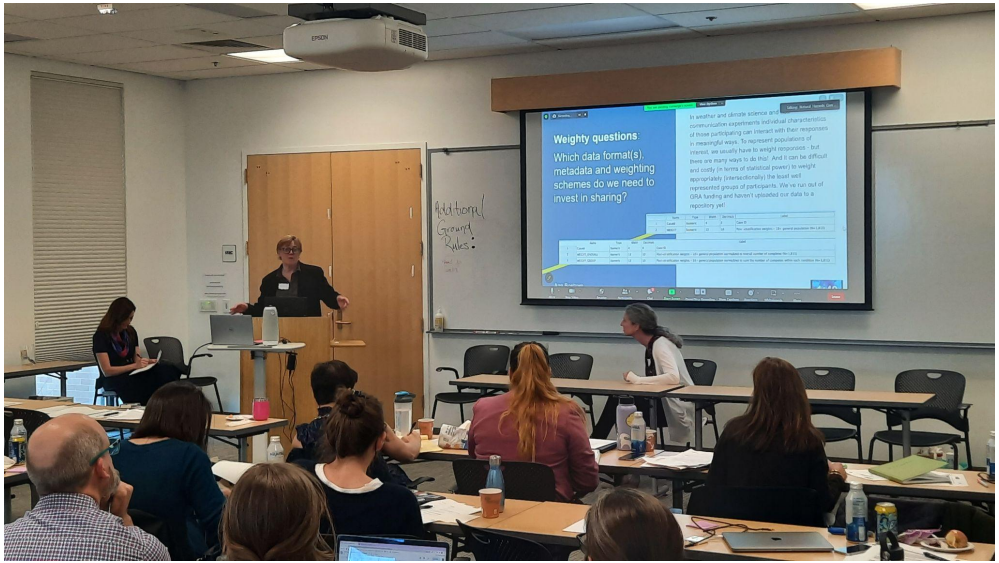
A robust data sensitivity and security policy helps foster trust with participants and stakeholders, as well as replicability of the final data collection methodology.

Critical infrastructure sectors are often siloed and so do not regularly share information.

Navigating their various needs, protocols, and comfort levels is incredibly challenging and requires significant effort.

www.RICHAMP.org

Part II: IRBs / Data Repositories / Metadata / Publications



How are Institutional Review Boards (IRB's) managing changing requirements for data publication? What if IRB guidance conflicts with funding requirements?

Jennifer Tobin

Institutional Review Boards across universities often have varying practices and protocols for what should be included in a proposal and how that is interpreted as ethical practice.

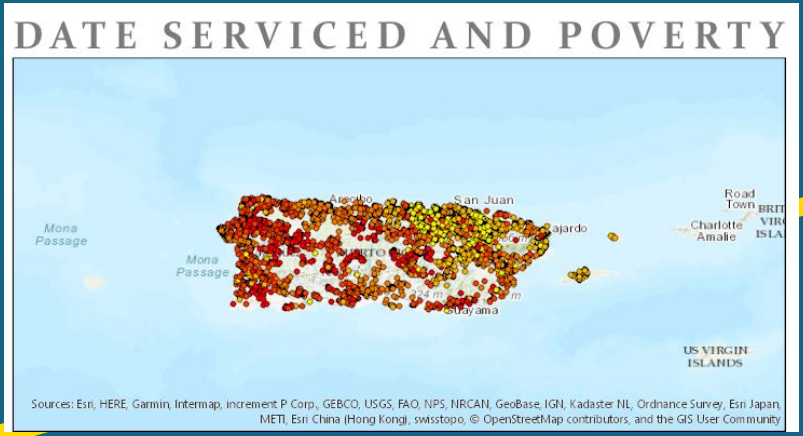
Is there any written guidance available yet that would help social science researchers get IRB approval to publish data?

Are there requirements as to which data repositories are used based on the home university of the researcher?

Do IRB's determine the level of de-identification necessary to publish data and have it reused?

NOTE: Qualitative Data Repository has been studying how IRBs are addressing new issues under the Common Rule. And IRBs have no good way to monitor what is happening with studies.

How do you decide *where* to publicly publish your data?



Mary Angelica Painter

My co-authors and I built a large original dataset on crew deployments for power restoration in post-Hurricane Maria Puerto Rico (which includes a PR specific social vulnerability index, poverty rates, and other variables). Without any experience or guidance on data publication, we were lost on where would best suit our data publication needs.

How do you incorporate several considerations when finding where to publish (size of data, type of data, language, etc.)?

Will the federal agencies that are requiring publication *require or recommend* that researchers use particular cyber-infrastructure?

How should (digital and analog) infrastructure supporting (qualitative) research workflows (including many stages of data selection and analysis) (involving tools/platforms like Atlas.ti, MAZQDA, NVivo, PECE, eEM Sketching) feed into (and interoperate with?) infrastructure primarily designed for research data preservation?

Kim Fortun

Related questions:

– Are research workflows themselves a type of data that should be preserved? If so, what with level of detail? Should only data content be preserved, or also the data structures used along the way? What kind of metadata is appropriate?

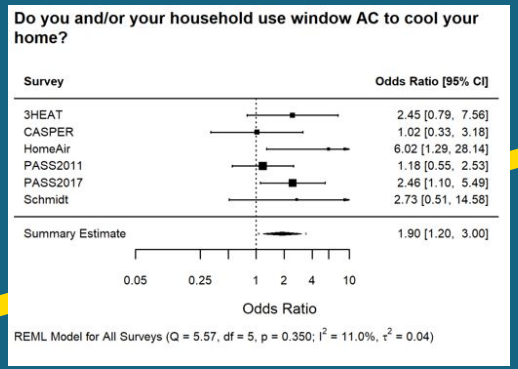
– How expansively should (ethnographic) data be curated, preserved and shared?

– Why preserve research data and workflows other than replicability?

– What workflow support collaborative analysis and hermeneutics?

– How can we teach the design and use of (digital infrastructure) (qualitative) research workflows?

If unpublished data are used in a meta-analysis or review, should they be published? By whom?



David Hondula

A student working on a publicly-funded research project performed a meta-analysis of previously collected survey responses.

Some survey data were available from public-facing repositories. Other unpublished data were provided directly from researchers. Some unpublished data were collected with public funding. (and some data that should be publicly available were not made available for the meta-analysis).

Who holds which responsibilities and rights to publish data sets used in the meta-analysis?

Weighty questions:

Which data format(s), metadata and weighting schemes do we need to invest in sharing?

In weather and climate science and risk communication experiments individual characteristics of those participating can interact with their responses in meaningful ways. To represent populations of interest, we usually have to weight responses - but there are many ways to do this! And it can be difficult and costly (in terms of statistical power) to weight appropriately (intersectionally) the least well represented groups of participants. We've run out of GRA funding and haven't uploaded our data to a repository yet!

	Name	Type	Width	Decimals	Label
1	CaseId	Numeric	4	0	Case ID
2	WEIGHT	Numeric	12	10	Post-stratification weights - 18+ general population (N=1,815)

	Name	Type	Width	Decimals	Label
1	CaseId	Numeric	4	0	Case ID
2	WEIGHT_OVERALL	Numeric	12	10	Post-stratification weights - 18+ general population normalized to overall number of completes (N=1,815)
3	WEIGHT_GROUP	Numeric	12	10	Post-stratification weights - 18+ general population normalized to sum the number of completes within each condition (N=1,815)

Ann Bostrom

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

- Some data posted online, such as tweets, are publicly available for research
    - Some users tweet about things that they would likely not want shared broadly years later (such as private concerns or illegal activities). Some also provide metadata (e.g., sequences of specific locations) that can lead to risk. When publishing, we go beyond human subjects guidelines to protect tweeters, e.g., by modifying, aggregating, or not using such data.
- If we share the original data sets, others may not follow the same ethical guidelines. Moreover, people may later protect their data or delete their account (or have their account blocked). Is their data still in the public domain?
    - We could share protocols for recollecting the same data, but what if the original platform no longer makes the data freely available?
- Related question: How do we handle data sets we are analyzing but did not collect, during this (long) transition?

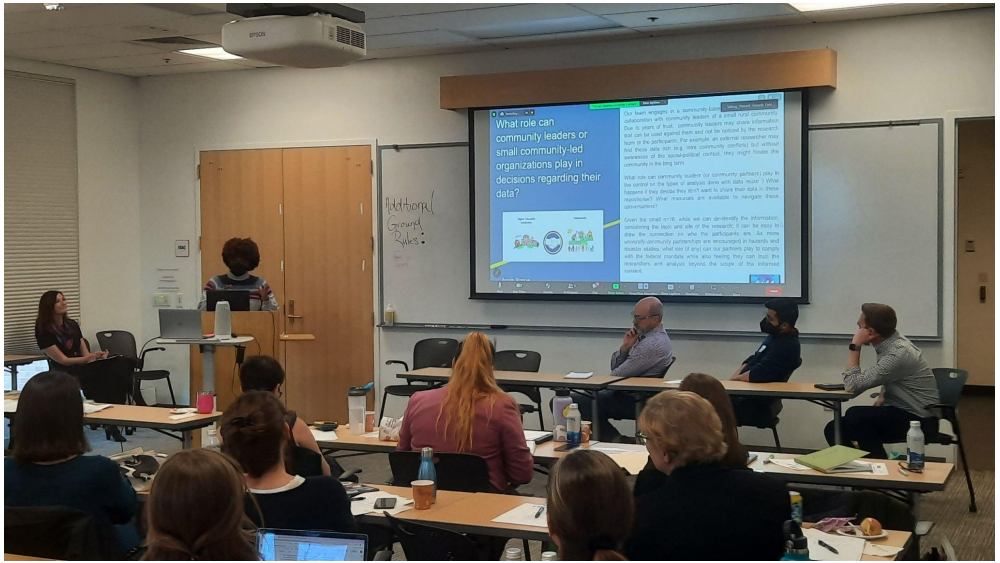
Rebecca Morss

How will the OSTP guidance be implemented in resource constrained environments and in diverse contexts?

Lori Peek

1. The OSTP guidance states that **scholarly publications** including a) peer-reviewed journal articles, b) peer-reviewed book chapters, c) editorials, and d) peer-reviewed conference proceedings should be made **freely available to ensure equitable access**. But many of these final outputs from research remain behind **paywalls**. How will this gap between vision and $ reality be addressed?
2. Researchers often **do not have the funds available** to support curating, publishing, and sharing their data (as most funds go toward original data collection). Will there be additional funding to help researchers adhere to new guidance? Or do we need to rethink how we budget funds and research time?
3. The OSTP guidance does not refer directly to the **publication of research protocols or instruments,** but this is a place where we have made progress as a social science research community. Will this be part of future NOAA or NSF guidance?
4. Colleagues in federal agencies must often go through additional peer-review for publications. When we have tried to publish together, we have run into issues where there aren't **established peer review processes within federal agencies for protocols, instruments, or data publications** in the same way there are for scholarly publications. Are federal agencies establishing review protocols?

Part III: Community, Data Ownership, and Data Reuse



What role can community leaders or small community-led organizations play in decisions regarding their data?

Anais Roque

Our team engages in a community-based participatory research collaboration with community leaders of a small rural community. Due to years of trust, community leaders may share information that can be used against them and not be noticed by the research team or the participants. For example, an external researcher may find these data rich (e.g. intra community conflicts) but without awareness of the social-political context, they might hinder the community in the long term.

What role can community leaders (or community partners) play in the control on the types of analysis done with data reuse ? What happens if they decide they don't want to share their data in these repositories? What resources are available to navigate these conversations?

Given the small n=16, while we can de-identify the information, considering the topic and site of the research; it can be easy to draw the connection on who the participants are. As more university-community partnerships are encouraged in hazards and disaster studies, what role (if any) can our partners play to comply with the federal mandate while also feeling they can trust the researchers and analysis beyond the scope of the informed consent.

The long journey from Community and Back Again…

Source: Wall Paper Flare

Nathanael Rosenheim

My Social Science Data Collection Timeline

2015 - Community partnership development
2016 - Grant Funding
2017 - Hurricane Harvey and RAPID Funding
2018 - Field Work
2019-2021 - Data cleaning and exploration
2021 - Data Publication
2022-2023 - Journal article preparation
202? - When will I get back to the Shire [Community]?

Frodo Baggins and the Ring of Power
(Photo: The Lord of the Rings New Line Cinema)

How do we design principles and agreements for co-authorship or credit upon releasing data for public use, and ensure compliance with our agreements?

Sameer H. Shah

Assuming our data is suitable for public release (e.g., quality assurance; de-identified, etc.), how should we best design principles and agreements associated with **data reuse** by other authors that ensures the knowledge contributions we made in creating the dataset are recognized beyond a simple citation? The time and energy invested into creating a data product should be recognized through co-authorship, or other suitable forms of input and recognition. Is it, on the flip side, appropriate to request this and to exercise a certain level of control and influence over how your public data may be used by others? What are some appropriate ways to navigate this, and how can we ensure compliance with agreements we create?

What do we do if people misinterpret or misuse our data?

Joe Ripberger

We began sharing datasets from the Extreme Weather and Society Survey on the Harvard Dataverse in 2020.

Our experience has been **overwhelmingly positive**. We've seen nearly 900 downloads and people are using the data for all sorts of amazing studies that we could not have imagined when we developed the surveys.

But once in a while, we notice that people misinterpret or accidently misuse the data in ways that make us uncomfortable. We are not sure what to do in these situations.

Can we do anything to prevent this from happening?

What do we do if we notice that it is happening?

How can we encourage "responsible" use of data?

Data collection and analysis require experience and expertise…