Developing an ESTS Open Data Policy
ESTS represented by Grant Otsuki, Aalok Khandekar, Angela Okune

In this presentation, we provide a brief overview of open data discussions in academia broadly, and the humanities and social sciences specifically, identifying significant tensions that need to be considered by STS researchers thinking about open data policies. Researchers are increasingly facing calls to move beyond publishing traditional research articles, and begin making their data available to the public. These movements are driven by funding agencies, disciplinary associations, and academic publishers, but also by emerging ethical norms demanding increased transparency in scientific research. Moves towards open data have been led by researchers in the natural sciences, engineering, and medicine, where open data practices are becoming normalized. However, researchers in the qualitative humanities and social sciences work under professional and disciplinary norms and standards that can make adopting the practices of the STEM fields challenging. In this presentation, we discuss past open data efforts in H&SS, and discuss key tensions that need to be considered in open data in the context of STS. We look at these issues from the standpoint of the journal *ESTS,* to consider and invite discussion of the role that an academic journal should play in
normalizing open data in STS.

Issues/Tensions:

- Establishing incentives for opening data. Making a persuasive case for open data.
- Platform standardization versus “ecology of platforms.” Establishing and maintaining platforms for open data.
    - Bibliodiversity?
- Discipline and interdisciplinary issues related to ethics of data.
- Training early career researchers and established researchers to think with open data.
- Making data discoverable, (interoperable), and usable.
- How to build flexibility into a policy document that should apply to all cases. (Nuanced universalism.)
- How to meet people where they are, but also raise awareness and encourage thoughtful engagement with open data issues.
- Prescriptive or inducing. We need a statement about this, this, and this. The Taylor and Francis aspects. “Ask authors to take a stand on whether authors will take reasonable requests for sharing their data.” Here are a bunch of options and we always encourage you to take the more open one. In your statement we need you to justify why *not* more open. We encourage creative and new interpretations of data and open.
- Some discussions for thinking about how people think about “open” and “data.”
- Check boxes to add:
    - “OPTIONAL: I would like to opt-in to participate in a pilot on open peer review processes.
    - OPTIONAL: I would like to be in conversation with ESTS regarding their open data policy and workflows.

Resources:
https://drive.google.com/drive/folders/1Z-JkmfQKO8eotW4fv1NFOR6Ja_w0dGRQ?usp=sharing

Expectations of the RA:
1. Build up the corpus of data for us to build our Open Data Policy
2. Analysis of that corpus and figure out the issues and tensions and where they manifest (see below for starting practical questions we have).

Questions and Areas of Analysis:

- Focus on qualitative data of the kind represented in *ESTS.* How do policies deal with the diversity and specificities of data and data practices? (e.g. of field notes, photographs, audio/video recordings, transcripts, survey instruments and results, primary sources/archival material, etc.)
- What are the major differences in open data policies? How can they be categorized? (e.g. prescriptive versus indicative; mandatory vs. optional; What are emerging or dominant trends?
- Are there any journals/policies that are frequently cited as best practices?
- What are seen as the key issues to be considered when publishing data (esp. In an academic journal?)
- In addition to the data artifacts themselves, what supplemental documents or materials (e.g. approvals, other governance documents) do we need to develop, publish or keep a record of?

End goal: ESTS Open Data Policy draft (and templates) to bring to our 4S roundtable in October as a proposal (of sorts). “This is what we’ve done, what we have come up with, what do roundtable participants see as the gaps? What issues have roundtable participants faced that we have not included?”

RA Milestone Deliverables:

1) Milestone: Google Doc of identified artifacts that can be uploaded to PECE (but not yet uploaded to PECE).
    a) First stab: end of July?
2) Milestone: Upload to STS Infrastructures, insert critical commentary for each artifact that summarizes the piece.
3) Milestone (if possible): Work with structured analytic for each artifact.

4 - 5 hours a week x 10 weeks. Ending around mid-Sept / Oct.

RA Activities

- Collecting open data policies from publishers, journals, scholarly societies, funding agencies, broadly with some focus on humanities and social sciences.
- Literature on Open Data policies, approaches, challenges, etc. (journal articles, blog posts, etc.) - see below for full list.
- Place all artifacts into a new designated PECE essay.

- Also create a PECE essay for IRB material guidance/templates that you may find along the way.

Suggestions for where to look for lit review:
Focused at the level of *policy,* starting with journals, but also at funding agencies, academic societies, data repositories, libraries/archives, and other actors as they pertain to the publication of data. Focus very practically: who has done a living will? How do publishers/authors distribute responsibility for caring for materials? (Including IRB stuff.)

- Digital Humanities Projects
    - Omeka
    - Scalar
    - QDR
    - Mukurtu
- Institutions/Centers:
    - MIT library (e.g. https://open-access.mit.edu/sites/default/files/OA-Final-Report.pdf)
    - California Digital Library / Calisphere
    - National Institute of Informatics (Japan) e.g. Japan Search https://jpsearch.go.jp
- Scholarly communities
    - RDA
    - Any other scholarly societies? (I know 4S does not but maybe sociologists? Psychologists? Prob more quant oriented...)
- Other Journals
    - Taylor & Francis
    - Wiley
    - Elsevier
    - STS specific
        i. Catalyst
        ii. Tapuya
        iii. STHV
        iv. EASTS - East Asian
        v. EASST - European
- Funding Agencies
    - Bill & Melinda Gates Foundation
    - European Research Council
    - Plan S
    - NSF
    - NIH
    - SSRC
    - SSHRC
    - JSPS
    - IDRC - explicit policy on OA
    - Wellcome Trust
    - Templeton Foundation

    - Wenner Gren Foundation (e.g. “Preserving the Anthropological Record.” Edited by Sydel Silverman and Nancy Parezo, 1992)

Longer Term OUTPUTS (for Angela and Grant with Prerna’s help):

- Guidelines for ethics review boards. Here are the ethics you want to think about when someone in the social sciences wants to open up their data. Something that others could use in an IRB application. - FOLLOW-UP WITH KIM on this.
- Initial open data policy, based on T&F and RDA momentums.
    - https://docs.google.com/document/d/1UWsZs4yXzQKIBBR-iDw27gfhPP0kxrkiJuy9KQ5ZAjM/edit#heading=h.mbjjtzu6nwvm
    - https://docs.google.com/document/d/1uV4-woWCK9kKRP4d26O2OecG5VrqFj_EM3hXaqK1p2g/edit

Meeting 22/23 July 2021

Table of Open Data resources, based on categories that were provided earlier.
Hyperlinked all of the artifacts, but not world through them yet.

Leading social science journals are from public health, and none are open access.
Biological/medical sciences have a lot of open access/data policies.
eLife: Public peer review, quite interesting. Also BioRxiv.
ASM and BMC also have detailed open data policies for human subjects.

Digital Humanities Projects
LOCKSS: Very detailed policies,
Hypothes.is: Annotations as a web layer. Everything can be annotated. Looks cool, but not sure how it works/useful. Partnership with Cambridge UP. And also related to QDR.

Reference Center for Environmental Information: lots of projects with different data types, but they don’t have a policy about how to archive or store.

More work needs to be done. In this section.

Scholarly Publishing and Academic Resources Coalition: Very cool comparison feature for data sharing requirements by federal agencies. Any categories to bring in from here to use in our comparison table? Lots of useful columns to bring in.

Funding Agencies.
Cool list of OA policies from Japanese institution, funding agencies. Japan Data Catalog for the Humanities and Social Sciences.

Interesting to look at different namings of policies. Eg. Wellcome Trust, “guidance.” Not so prescriptive.

Journalism: The Guardian. ProPublica “Data Store”: Different kind of data sets. Interesting move among journalists to publish raw data.
Also Bureau of Investigate Journalism. Bureau Local. Connect local level stories and the raw data behind it.

Getting in touch with QDR for a discussion?

Harvard Dataverse. Widely used. Similar to OSF? (Skewed towards quantitative rather than qualitative.)

Non-English journals. Journals/publishers outside of the US and Europe?

Citational practices about the digital humanities practices do. In comparison with journalists’ practices. What are the practical steps to engage people with the data?

Muckrock: Millions of artifacts. Freedom of Information Access space. Very deep repository of data, mostly in the form of PDFs. Tools, nested services. Heavy technical learning curve to the process. Steve Jobs’ CIA record.. Using FOIA requests to get data and make it public.

For the end of July:

- Adding a few new sources.
- International publishers.
- Digital Humanities projects.

- Looking at the language.

Check in on Slack at end of July.
Meeting a week later.

Meeting 9/10 August 2021

ESTS Open Data Policy Analysis:

Using SPARC comparison tool to develop categorizations, using questions also from the OA/OD analytic from STS-I.

- Inclusions and exclusions of data types. Funding agencies defining specifically what they want to include and exclude. All want to exclude peer review and communication with colleagues.

- Also, would be useful to find sample documents that could be used as a basis for our own policies.
- DoE has glossary to explain all of their terms.
- A: Questions from style guide. Citing data, and uploaded their artifacts into STS-I. Using MIME-types to categorize different file types as a typology for artifacts.
- Need to write-up a workflow to show the quotidian steps for working with Open Data.
    - “How do users move through the system?” is not outlined in many policies, but would be useful to have.
- Reading for metadata (bottom of page 4): What metadata is required, needs to be included?
- P: Next, when done with what data types are included/excluded, then will move on to metadata. Will make different Google docs for each of these sections.
- NSF: broad general policy and then there’s a specific one for social and behavioral directorate. One suggestion would be to focus on specifics to social science and qualitative data.
- NSF asks for a specific DMP: give people opportunities to use existing DMPs or provide different pathways? The NSF one specifies where things are archived and preserved, for example.
- Which/when artifacts to upload to PECE? Many policies are derivative of Fair Use or EU guidelines, so would make sense focus on those “origin” policies.
- Questions for people on the 4S panel to address, and present from our standpoint on the prompt questions, and bring others in.
    - Particularly questions, within in the analytic set that you see as particularly generative?
        - Who was the system built to serve and why?
        - Who built the system and with what social, political and economic commitments?
        - What assumptions about language and knowledge are built in?
        - Policies often exclude peer review and peer communication. A question that addresses the assumptions about collaboration in the policy. How that collaboration is made visible or not visible.
- Not only opening data, but also gatekeeping data. Needing documentation to access that.
- Difference between opening and sharing data: as part of an archiving/preservation service, they do data cleaning protocols without having it open: access control services
- What guidelines do people use to archive and preservation protocols?
- In terms of reusability etc, what's the difference between opening and sharing data?
- Open data: gradation and modular data
- Linking data modularly that is already publicly accessible
- Would photograph of a copy-right material being uploaded on PECE be a violation of copy-right? After metadata, go into IP
- Open data policies could suggest not doing copyright and leave it to authors
- Policy about inclusion of images, do they require explicit consent from the creator?

- Ask authors to include an open data statement. Esp. if their material is copyright. Explaining why it's not accessible or can't be posted
- Interactive quiz kind of setup for getting people to think about already existing data practices and getting at what ESTS wants

Meeting 23/24 August 2021

- Documents: ESTS Open Data Policy Analysis Main

- Open Data Policy Research Table: list of policy documents being compiled for later annotation

- ESTS Open Data Policy Analysis: Inclusions & Exclusions of Data Types

  -including a glossary/FAQs clarifying vocabulary of open access/open data
  -most funding agencies exclude preliminary material as research data
  -Sage uses QDR for identifying what constitutes qualitative data. QDR further has "data project types" that aid authors in depositing their data. The type "ATI Data Supplement" can be useful for PECE-style annotation/analytic process. Requires a Data overview Statement of 1,000 words that asks authors to think about why they chose particular data/datasets/evidence, instrumentation, and omissions/possibilities; why they made analytic choices; and "logic of annotation" i.e. why a particular passage was highlighted for annotation, etc. Sample ATI project

- ESTS Open Data Policy Analysis: Data Sharing & Management

  QDR policy for working on sensitive research data: "We discovered that most IRBs had not yet begun to carefully consider how their regulations interact with data sharing, and that few IRBs offered researchers concrete guidance on sharing human participants data, or templates for informed consent scripts that anticipate data sharing." They offer Guidance for IRBs to aid this

  Wiley: encourages, expects and mandates
  Taylor and Francis: table comparing different data sharing policies

- ESTS Open Data Policy Analysis: Metadata and Documentation

  -data availability statements: restrictions of repositories, third-party restrictions. Wiley and Taylor & Francis both leave it up to the author to decide which repository to use and what data availability statement they should submit. Wiley asks authors to "Visit re3data.org or fairsharing.org to help identify registered and certified data repositories relevant to your subject area".

-data notes by T&F: peer-reviewed article type description of repository use and analysis. F1000 has a detailed template for data notes
-COPE guidelines on Image Manipulation, suspicion of fabricated data: apologies and responses flowchart
-F1000 Research platform has examples of metadata and data citation practices:

This article includes interviews (labelled qualitative) with a section on consent and a data availability statement: "When using the findings and data of this study, the small and selective sample should be born in mind; these give an overview of the interactions experienced by this particular group of researchers and the ability to extrapolate to the general population of researchers may be limited."
Data citation: "Burgess, Heather; Chataway, Joanna (2021): Research collaborations and research meetings in African Health Research - Questionnaire data. figshare. Dataset. https://doi.org/10.6084/m9.figshare.13726087.v1"

This article also includes interviews and provides links to "extended data" in the Mendeley Data platform which includes a sample of hand-drawn diagrams (life timelines) of participants, a workshop evaluation form, sample consent form as an appendices
Data citation: Meherali, Salima (2021), "Girls Empowerment Project_FGD Guide", Mendeley Data, V1, doi: 10.17632/tp585h7nsj.1

- ESTS Open Data Policy Analysis: Citation and Attribution

Repatriation of artifacts and indigenous data sovereignty
Data as relationality, CLEAR, lab book
Mukurtu CMS
Generating meta-value statements: Tension between open and shared
Sharing data is better or "goodness" of sharing data: Moving away from data as commodified and not just data as object
Next level: to realise a data relation, wherever data can be shared, it should be shared
Then about tensions, ethical considerations (when is it ethical)
ML "i am fighting science with science, not indigenous science"
Active refusal

Next steps:

1. Reading briefly about restitution, repatriation, refusal (just a few hours)
2.

(1) reading OA-OD policies from scholarly communities
(2) repatriation of artifacts both physical and digital
(3) restitution of artifacts
(4) 3-D copies of museum artifacts: