Data Protection and Sharing in tension

Thinking about future developments

Initiatives to protect data

- Considerable research potential for data that has a protected component
    - Social science data
    - Electronic Health Record data
    - Biometric data that retains identifiability (fingerprints, facial scans, DNA)
- GDPR in Europe - strong protections, high standards for security
    - US Cloud providers aren’t certified
    - Processing of biometric data that identifies a person is prohibited (except for the deceased) – GDPR Art 9
- In US rules for health (HIPAA), education (FERPA), financial (Gramm-Leach-Billey act)
    - Research data falls under Cybersecurity Maturity Model Certification (CMMC), HIPAA, NIST 800-171/53
    - NSPM-33 which starts in 2024

In tension: Open Data Standards

- Open Research Europe under Horizon Europe grants
  - Publish data for provenance, re-use, and further research
- US NIH Data Management and Sharing Policy
  - Likely to be followed by sharing policies at NSF, others
- FAIR data guidelines

When to protect? When to share?

- Current rule of thumb is at publication, until 6 years after last citation
  - FAIR provides guidance about how it should work but not how to make it sustainable
- Real questions about where things will be shared, with whom, and what expectations are?
  - Who tracks the citation and determines if things are shared adequately?

Another wrinkle: Research Reproducibility

- For some institutions, this becomes a source of pain
    - Targeted reproducibility of politically or otherwise sensitive research
    - Retention now stretches out considerably
- For some institutions, this is an opportunity
    - Data can be re-used for new studies
    - Reproducing studies is an opportunity for young scholars to engage with the research stream

Participants

- Funding agencies
- Researchers
- Research Leadership and IT
- InfoSec, Privacy Officers, Counsel
- Libraries/Archives
- Antagonists

What does the future look like?

- Higher security requirements across the board are likely
  - Even for basic research, nations are concerned with IP theft and research scoops
  - Data protection means that international collaborations can be fraught
- NIH is providing data-sharing platforms in the cloud but it doesn’t seem that other agencies have a motivation to do this
- Increasing data scale/resolution/complexity means that sustainability is constantly getting harder
- Individual disciplines probably benefit the most from large scale sharing initiatives
  - That means that interdisciplinary collaborations have to break down silos to be effective

More visioning

- Adversarial model
  - Sharing doesn’t necessarily mean “free”
  - Institutions may begin treating data as an asset to be protected instead of a headache to be retained
  - Limit the number of reproducibility challenges and create a flow of funds to sustain retention/protection
  - Does this create a data market? Are there risks of clientism?
- Collaborative model
  - Discoverability as opposed to FAIR’s “Findable”
  - Federated datasets that allow for interdisciplinary collaboration
  - Federation requires considerable agreement between parties
  - Sustainability remains a question