GDCC Sensitive Data IG Meeting #2

Date

2022-11-28

Participants:

Name, Affiliation, Email
Cheryl A. Thompson, UNC Odum Institute, cathompson@unc.edu
Don Sizemore, UNC RDMC, dls@email.unc.edu
Sonia Barbosa, The Dataverse Project, sbarbosa@g.harvard.edu
Amber Leahey, Borealis amber.leahey@utoronto.ca
Steve Baroti, University of Toronto Libraries steve.baroti@utoronto.ca
Katie Mika, Harvard Library & IQSS, katherine_mika@harvard.edu
Jonathan Crabtree, UNC RDMC, Jonathan_Crabtree@unc.edu
Betsy Gunia, Johns Hopkins University, betsy.gunia@jhu.edu
Bob Treacy, IQSS, rtreacy@g.harvard.edu
Dave Fearon, Johns Hopkins University, dfearon@jhu.edu
Sebastian Karcher, QDR, skarcher@syr.edu
Leslie Barnes, University of Toronto Libraries leslie.barnes@utoronto.ca
Jim Myers, GDCC, QDR qqmyers@hotmail.com

Agenda

1. Welcome and Introductions
2. Current Capabilities and Workflows for Sensitive Data in Dataverse
    a. Sebastian Karcher, Qualitative Data Repository (QDR)
    b. Sonia Barbosa, Harvard Dataverse
    c. Jon Crabtree, UNC Research Data Management Core
    d. Jim Myers, Global Dataverse Community Consortium and QDR
3. Discussion
    a. Discuss sensitive data needs, gaps, and desired features
    b. Use cases for sensitive data
4. Administrative
    a. Next topic
    b. Meeting frequency

i. Once every other month
ii. Alternate meetings between time zones?

Minutes & Notes

- QDR (Sebastian) - Data repository for primarily qualitative data
  - Repository DOES contain sensitive and potentially identifiable data
  - Fully curated, partly because of the risks sensitive data carries
    - Review consent forms regarding the sharing of sensitive data
    - Researchers must be involved in deidentification process, largely due to specific contexts, indirect identifiers, and local knowledge
    - Offers de-identification guidance & disclosure risk reviews. For every transcript in a project, and samples from very large projects (100+ transcripts)
  - Sensitive data matrix: https://qdr.syr.edu/policies/sensitive-data
    - This is framed as a conversation with the researchers
    - Not formalized due to risk of promoting “security theater” rather than precise evaluation of data sensitivity
  - Access controls developed in dialogue with researchers. Who may obtain access to data? Manageability is important, customized for each project. “As open as possible, as closed as necessary”
  - Technically
    - Encrypted at rest in addition to at transfer
    - AWS standard encryption, Storj (distributed/decentralized storage)
    - One dataset is too sensitive for Dataverse storage. For this one, zero byte data files are uploaded to Dataverse along with documentation and supplementary files. Data are stored securely on Syracuse U servers, pw protected, encrypted zip files are sent. PW is given over zoom call.
    - https://data.qdr.syr.edu/dataset.xhtml?persistentId=doi:10.5064/F6V5VGX3
    - Keycloak & OIDC for authentication
  - Generally low-level access controls with minor DUAs - no institutional signatures, but researchers are vetted by QDR team via application & agreements to destroy data at project conclusion
    - Handled via https://qdr.syr.edu/discover/specialdownload (requires QDR login)
  - Researchers are generally involved in approving access, with some fallback protections if researcher become uncontactable (QDR team with review board)
  - Support customized guestbooks for requesting information from data users for facilitating data access; most file requests are also followed up by email and in some cases through Zoom (currently) for authentication
    - Guestbook entries are reviewed before access is granted.
  - Guestbook at file? request is in the upcoming Dataverse 6.1 release

- QDR has a policy about removing data and revoking access when required, but difficult to determine, and managing in DV is fairly straightforward. Versioned files/datasets must also be managed accordingly in those circumstances.
    - Others have had to deal with this, where some subjects needed to be removed from the data files. Previous versions couldn't be controlled for, and beyond removing records, restricting the data after the fact was another solution applied.

Harvard Dataverse (Sonia)

- Does not accept sensitive data
- Murray Research Archive - all de-identified data, some contain audio/video that are sensitive and are not kept in Harvard Dataverse
    - Data Access Forms - Qualtrics hosted, very long, reviewed against restrictions on the data
    - Restrictions are reviewed / approved by researchers and archive admins
        - As people move on, this is transferred to the archive
        - Copyrighted measures ; still managed by archive
    - Terms of Use - clearly states DO NOT DEPOSIT Sensitive data; some are not aware, others take this very seriously; researchers are responsible, Harvard is not liable - to some extent
    - All datasets that are restricted have conditions of access ; e.g. clinicians present for reuse, etc.
    - De-identification process - review MOU for deposit, review cleaning initiatives for EACH dataset, PI inform the archive about the sensitivity of the data, if any questions some campus partners (ethics/security) are brought in to determine
        - Thorough review of variables
            - Aggregate data (birth dates, etc.), remove data (PI data), also always keep an original copy (as a backup)
            - Keep a record of everything that is removed
            - Handbook of de-identification
                - Helpful for qualitative and quantitative data
        - De-identified data is what gets published
        - Does not cover audio/video materials , only hosted on-site, and additional controls / security are applied
        - Few researchers accidently share personal identifiable data
    - Guestbooks - to capture who is using the data (sometimes this is all that is required for use)
    - Murray examples from Sonia:
        - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4POMAK: Sensitive, will update restrictions to include requires mentorship with licenses clinical provider based on type of data used and conversations with the primary user of the data
        - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CR02PG: This dataset had variation in their restrictions per subject

inserted into the file level description. Subjects gave approval for their names to be used. Was restricted but recently changed with permission from PI, to utilize just the guestbook.
        - Harvard Repository collection, used Murray Terms for their data but not Murray affiliated: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XPI0OW

- Example Texas Adoption Project Dataverse (harvard.edu) with many conditions for use, each file has different restrictions / conditions for use
- Sometimes when the researchers move on, we can't change the conditions
- NIH GREI and OpenDP Project (https://opendp.org/)
    - Seems to be the best tool to start with for implementing in Harvard DV
        - Requires permissions from Harvard for implementing / enabling (they are open to researchers at Harvard using this, and some affiliates, but challenging to implement for other non-Harvard researchers to use)
    - Tool for generating differentially private statistics for sensitive data. Give access to summary statistics of sensitive data with some added noise to reduce identifiability

UNC - TRSA/Impact Project (Jon and Don)

- UNC proposal approved last year for RDM core group (storing / deposit of sensitive data in DV for all research departments/groups across campus) – ramping up, hiring, and ODUM staff will be supporting
- Current UNC Dataverse is self-deposit that does not accept sensitive data
    - Clear gap for sensitive data
- Investigating an integration between Local storage for sensitive data that can be published to DV
- Going to be leveraging multiple systems,
- For FDA studies with sensitivity HIPAA data and require rules and regulations around access / storage
- Created a cloud instance using AWS that is validated and HIPAA compliant closed storage only for authorized users (clinicians, etc.) and integrating this with DV
- Looking into expanding/adapting this cloud instance for others, for a broader group of researchers
- Working on a grant project to automate DUA to access secure storage with Dataverse in the front-end ; goals: store sensitive data in secure cloud storage, approvals can be more automated

Dataverse (general features) (Jim)
Design Considerations for Handling Sensitive Data in Dataverse - Google Docs (PLEASE CONSIDER REVIEW / ADDITIONS)

- Addresses where are we now and what are the technology gaps?
- US regulations for storing sensitive data; software can help but it's really about processes around the software

- Provides an overview of how DataTags COULD work with these features to support classification / management across the repository
    - DataTags adaptability for institutions
    - Designed as a questionnaire for assigning appropriate tags to data
    - Could be leveraged by an external tool
    - In an ideal world: Once tags are assigned then the system can do certain things (MFA, storage and encryption, DUA signature, and controls applied, restrictions, etc.)
    - But overall, most of things are supported by DV already, but not currently tied to workflows/controls for sensitive data
    - Looking to develop more of the system interactions so this kind of management (submit for review, access controls, etc.) can be made more streamlined
- Encryption
    - Mechanisms such as encryption by depositor
    - Storage encryption
    - Transfer encryption
- Remote storage
    - Remote storage linking (provide a URL to file or front-page of another system) is now supported in DV

Action items

- Goal for group - sharing information about internal policies for managing sensitive data (de-identification, determining access controls, etc.)

IG History and Related Documentation

2023-09
GDCC Sensitive Data IG Kickoff Meeting Notes
https://docs.google.com/document/d/1Gv2uHWqTrDME7WFwMYPjDmeDwnBsLXiO1lZ_iAR9lgE/edit?usp=sharing

2023-06
Dataverse Community Meeting discussion
https://docs.google.com/document/d/1vFz0HShR33QKKomdXIF90uRLb5bv8-dxe2cco8ckiBU/edit#heading=h.g8fnwr7qvnao

2023-01

Draft whitepaper

Design Considerations for Handling Sensitive Data in Dataverse