Shared Notes for the In-Person Meeting for the IMLS DDP Storage of Private and Sensitive Data Grant

December 6 2019, Austin, TX Agenda
Link to slides for the day

Thank you, in-person and remote attendees!

Welcome and Introductions

Name, institution, interest in the project

Presentation: Overview of the Investigation

Definition of DDP:

- Importance of geographical distribution across regions, avoiding overlapping disaster zones, power grids, etc.
- Nodes under control of different administrators.

- Content at sites should be on live media that can be checked constantly.
- Network should perform three main tasks:
    - Ingest or harvest content
    - Content monitoring
    - Ability to retrieve content
- Also reference to the Digital Preservation Storage Criteria (Sibyl)
- Jaime: in theory DDP network as defined above is great, but in practice is difficult. Especially at scale. How do you handle access to this content?
- Courtney: Difficult to assess how much content is really out there because the need for DDP exists before local thinking/readiness has happened.
- Accessibility of the archives in these networks.
    - Chip: Our current DDP services are dark archives and not set up for access.
    - DuraCloud plays a role here, providing easier access to materials in storage.
    - Identify management in the distributed network vs readily accessible data for depositors in that network
- Assumption of the project: if we can meet HIPAA standards, we can meet standards for "everything else." An assumption that may be challenged by this investigation.
-

Problems

- There is no DDP service dedicated to preserving sensitive data
- HIPAA requires that certain data be preserved

State of preservation at health science centers currently. Are there local preservation systems?

- UTSW has data center in Arlington where everything is stored (a UT System service)
- Johns Hopkins - not one that is good for everything

Role of cloud providers

- Problem with lack of transparency in AWS integrity checking
- Shared AWS service model allows for you to use tooling to audit the content you want to audit
- APTrust uses AWS storage and services
- DuraCloud and TDL also offer storage options in AWS
- Concern that so much cultural heritage data is getting collected into ONE commercial service provider (i.e. AWS)
- Chris Jordan: this is an area where commercial and tech changes really quickly. SIze of the academic market is very small comparatively. Willingness of cloud providers to work with small scale institutions may change overnight.
- APTrust tries to mitigate risk by adding other commercial storage providers, not just AWS.

Components of HIPAA compliance (Jen) - primary components are confidentiality, integrity, and availability

Q&A

Notes

Small Groups Use Case Report Out: Use Cases for Private and Sensitive Data Preservation

Group 1 notes:
David B: The Benson/UT have a strong collecting focus in human rights documents, particularly in Guatemalan National Police Archives (AHPN). This is a physical repository in the country that is the records of the Police - lots of records documenting human rights violations from 1960-1996. Exposed in 2005 and made available and protected via offsite storage. The current government wants to shut down access to this material and there are questions about UT's role with this material. It has been written to tape according to the standard UTL digital preservation approach. There is another collection in Guatemala that they are looking at, produced by the Office of the Archbishop. This collection contains human rights records from the Archbishop's Office's alternative investigation into human rights abuses produced during the postwar peace process, and may be under threat from the Guatemalan government. There is a digital archive in existence but there are issues with making sure that it's safe and protected. UT's interest is to make sure that the needs for protecting the content are being met, especially when their initial reaction is that it shouldn't ever be distributed about multiple nodes.

Lyrasis - we don't have datasets in the same sense that universities do. But I work with customers within DuraCloud who do have needs. At this point because there's no sensitive data capacity, we can only take it if they do the encryption the data locally before they hand it off. We've talked with enough folks that have said they would use this, so it would be important for the community to have something like this in place. One issue is that much of what goes into our system isn't well-curated.

Defining sensitive data: Is it up to the collection owners to decide whether information is sensitive and needs the highest level of protection? Should users choose the level of security (and cost) based on the data they are submitting?

UCSD: All the data that goes into the UCSD repository is curated, controlled and described. The repository managers have a pretty good idea of everything that's going into the repository. The medical school is a large part of UCSD, but the repository stays away from handling any records that would require HIPAA compliance. There are some occasional grey areas (doctors' teaching records, e.g.). But the repository managers are interested in being able to offer HIPAA compliant storage. It's definitely a need that many users talk about, and we'd like to be part of the solution rather than just ceding it to the commercial space.

At Texas State, we would like to get better at working with unprocessed materials. Once they are processed or accessed, it may be that there are strict rules or agreements in place that we need to live by. Or they may be complete garbage or not worrisome. So there are definitely times when we really don't know what is in the collection when we start working with them. We have many examples of things that present challenges as to whether it's sensitive or not, for instance video or audio.

Group 2 notes:
Challenges:

- Data that is not de-identified. Trying to figure out how to de-identify the coordinates using encryption. Permitted to share internally, but research should be abstracted.
- Human rights documentation, government abuses, politically sensitive. Currently kept offline on tape, not preserved.
- Unprocessed collections, even appearing in unexpected place - architectural collection, sifting through data, transforming. "Unclassified data"
- National hazards engineering data - data that logs coordinates of house damage, photographic evidence of housing disasters.
- Separating out the local access part
- Certain known states- actively worked with or not. Ownership - who owns and manages the information. Manager implements controls around decisions on who can use. Ownership of data being driven by commerce (US) vs privacy of individual (EU). Without knowing who owns the data, hard to make good decision. Data owners can make decisions unless the law says other things.
- Data encryption - key management service, tension between digital preservation and encryption. Role vs person as the data manager. Also role of the DDP.
- Who owns patient visit data? Doctor? Patient? Insurance?
- Destruction of data is the ultimate ownership act of data- who is allowed to destroy
- Legal deeds of gift - do they adequately handle digital materials? Deleted files on a hard drive.
- Research studies- does the participant know it's going to end up in a long term archive?

Group 3 (Chris, Jaime, Kelly, Chip):
First Breakout notes:

- Cloud services: Some encryption risks not addressed
- Changes in institutional infrastructure--much moving to the cloud, so how long-lasting will be institutionally-based answers (existing and developed in the short-term future)?
- Medical data warehousing--is there a use case somewhere in this emerging practice?
- Evolving legal environment: data preserved at a single point in time in that evolution. What are the implications for DDPs when data was received under one expectation of privacy that changes as the legal environment does over time?

- Do the use cases involve data/materials that need to be accessed, or just disaster planning?
- What kind of access is actually needed in long-term preservation of sensitive and private data?
- Jaime: all projects have beginning and end, but that doesn't mean that interested folks are finished with uses of the research products.
- Chris: patients, researchers different perspectives on privacy
- Kelly: new interest in old written records that are being digitized; patients seeking family history information vs persons who want to protect family history
- New research interests seeking access to old seemingly unrelated data that may contain sensitive/private data.
- Ownership of data
- People depositing unprocessed material that they intend to process at a later date
    - Would DDP depositors want better, more secure services?
    - Would DDPs like to take more responsibility for those or share that functionality via a speciality service (and operated by who)?

Group 5 (Ramona, Chianta, Susan):
New use case of linguistic and language documentation data:

- Informed consent for participation in research - consent can be revoked at a later date, but researchers put still put personal names or initials of participants in the digital filenames
- Sensitive info might be hidden in A/V recordings that are in minority Indigenous languages that no one who works at the archive speaks - such file names can be a challenge for preservation
- Research data that includes personal narratives about traumatic experiences narrated by women in a part of the world where women are considered to be property; the very fact that these women are telling these stories and allowing them to be recorded (audio/video) puts them in a dangerous and vulnerable situation with respect to their communities. These recordings/data need to be preserved, but the women's identities need to protected, in some cases the stories need to be suppressed so that women cannot be (mis-)identified.

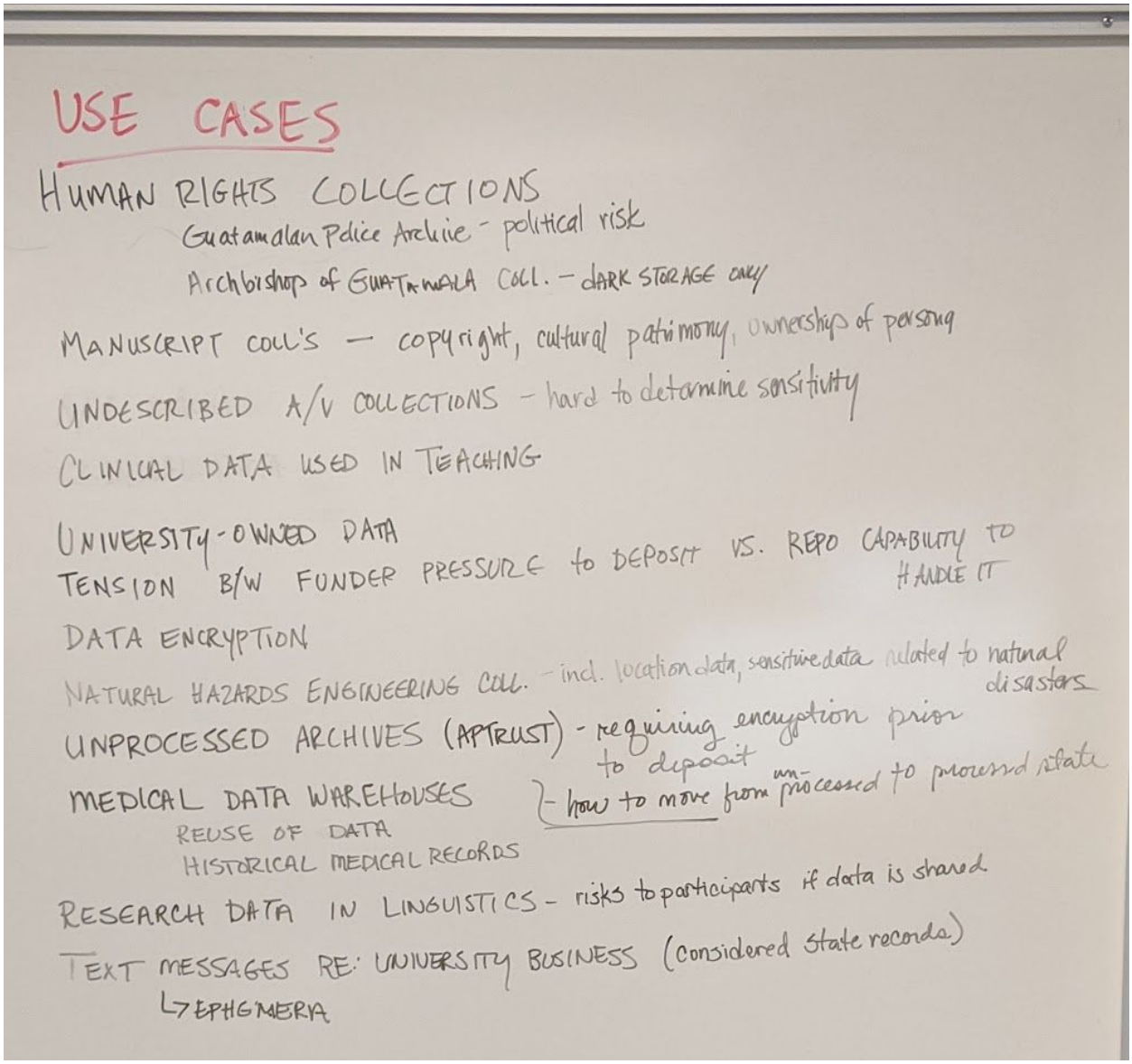
Medical archives, hidden PP/HI in archives:

- Boxes of medical records from the 1990s found in a closet; access to the closet no longer controlled (b/c no one knew that PPI, PHI was stored there).
- Mixed bags of physical archives: PHI mixed in with non-PP/HI documents that get sent to the archives for scanning; boxes not properly labeled.
- Educating campus contributors about handling of sensitive materials is a challenge, they know what hipaa and ferpa are yet more work needs to be done to ensure data is secured.

Report out / discussion of use case groups

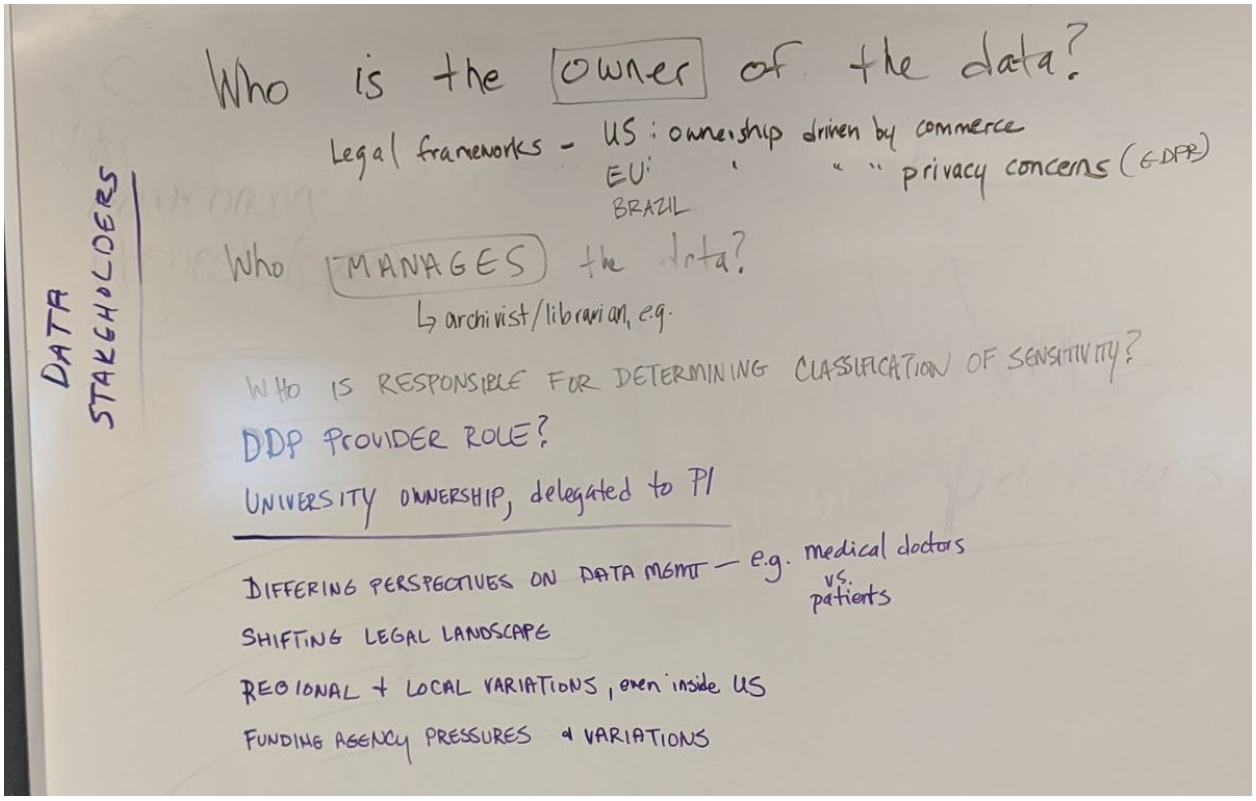
WHITEBOARD PHOTOS

Use cases:



Another use case: Potentially patentable data. Universities sometimes designate data according to its patentability (potential for commercialization, etc.).

Data Stakeholders:



Data ownership (Jen)

- In any use case, management decisions come down to data ownership - owner decides ultimately about destruction of data; decisions about use come down to ownership, too. Drives policies and management decisions.
- In US, ownership of data is driven by commerce; in EU, it is driven by privacy.
    - California and NY are exceptions to US commerce driver
    - Capitalism is a significant impact on this project and on private and sensitive data in libraries
    - Archivists/librarians - data custodians who decide how and when the owner decisions are applied; classification and regulations apply
- Group 1
    - David B -
        - Guatemala - digitized police records from civil war there; UT has been an offsite digital backup of this archive since 2010; public-facing but Guat gov is asking for return and removal of the website; offsite distributed copies so that Guat cannot tamper or easily demand the return of the content
            - Guat as data owner complicates this issue
        - Not yet acquired: totally dark storage, human rights collection from Guatemala; government might confiscate; Office of Archbishop of Guat- parallel collections relating to the UN peace process

            - Have to prove to partner that it is still getting preservation treatment and secure when distributed
    - Lauren G -
        - Manuscript repository - copyright issues (small piece); what kind of ownership does a ‘creative work’ entail; cultural patrimony; unprocessed AV materials (especially problematic because of the large backlog, could be undescribed and unclear what might be in it); AV also involves multiple parties (representation of a persona, including Screen Actor’s Guild)
    - Bill B -
        - V. difficult in AV especially to determine the security and processing needs as well as ownership and rights issues
    - David M -
        - UCSD Medical campus - grey area where teachers are mandated to store and preserve; liability concerns create inaction about storing it
            - Responsibility for assigning classification / who is responsible for determining whether the content is sensitive or not; Is it the donor responsibility to identify or is it the institution’s representative at the cultural heritage institution
                - In a lot of cases, we abstract the liability for private and sensitive data - push to the content depositors
                - DDP networks assume that the depositors have the expertise to identify sensitive data; most DDPs don’t check for the sensitive data
            - Johns Hopkins owns the data, but the PI has responsibility for managing the data over time
            - Susan K - often a federal funder requires that the data be publicly available; pushes the PI to put everything in a repository without deep understanding of what the repository is for and the repository manager is challenged by their ability to handle all the data the PIs want to dump
    - Group 2
        - Shi - Encryption and transfer of data over a network
        - Ashley - Field recon/research - catalog homes and coordinates after a disaster; people hadn’t consented to having their possessions and homes photographed
    - Group 3
        - Chris J - Need a broader notion of data stakeholders to encompass all the use cases represented here; who has a stake in defining the privacy and the sensitivity of that data and of its value into the future
            - Unprocessed archival submissions
                - Donors handing over entire corpus of their work/life, a lot to weed through

        - At APTrust, depositors do the encryption and manage the keys before it's deposited into the network; is it being audited over time to allow for change in policy/regulations
        - Accessibility - instead of encryption, can it be only a local medium and in a secure cabinet; avoids problems of key escrow, etc, but fails at distribution and other aspects of good DDP
        - Cold storage - issue of moving into an accessible place once it's been processed/assessed/appraised/cataloged
    - Would archivists use a DDP network for private and sensitive data if it was available - esp before all these issues of ownership and stakeholders are not settled
    - Medical data warehouses - conflicting interest esp of older med records
        - Different actors have different perspectives about how data should be treated
        - What is PII today will change over time.
- Group 4
    - Ramona and Susan - Narrative work from a part of the world where the participant could be at risk if the information was revealed
        - NSF is asking for public sharing of data in an archive or repository, and the researcher wants it to be preserved because of the risk of women's speech in an indigenous language; as a female researcher, she had access to male researchers can't go. Safety of the women she worked with is at risk

Other issues:

- Texas Senate Bill 944 - causing campuses problems; text messaging is considered a state record and business of the university; mandate to preserve
    - All transitory records are required to be preserved
    - Nathaniel - screenshots!
- UC records are all considered state records
- Chip - Laws that govern different nodes in different jurisdictions
- Susan - European data, Brazilian data laws
- Chris - solicited donated materials from events that contains sensitive data
- Ramona - Destruction policies and procedures

Current state of agreements in DDP networks and private and sensitive data storage services

Chronopolis storage at UCSD is library-owned storage resources co-located at SDSC. Library manages resources, networking, etc. (Need to update diagram to reflect that UCSD Chronopolis storage is UCSDL storage @SDSC).

Asymmetry between deposit agreements of TDL and UCSD

“You can’t write an MOU to wake up in the morning that would satisfy the needs of all institutions.”

Clarification re: DuraCloud

- Lyrasis manages a DuraCloud instance
- TDL manages its own DuraCloud instance

Bridges both live at SDSC
Data centers -- TACC, UCSD/SDSC, NCAR, UMIACS -- are geographically dispersed, managed separately, use different models.

What’s missing in the set of agreements? What roadblocks exist to contracting at your institution?

- Local readiness is a roadblock. Context: demise of DPN.
- Lack of vision beyond the institution. Difficult to look beyond what kind of data is handled on the campus and understanding that universities hold more than student records.
- Different institutions have different levels of maturity in terms of assessing risk of their own materials.
    - https://uit.stanford.edu/guide/riskclassifications (example of a mature model)
    - Most institutions do not currently have digital preservation policies for private and sensitive in place. Do they need to have that as a precursor to any service. Secondarily, can we write institutional policy that doesn’t get overridden by funder requirements, etc.?
- Different levels of requirements at different institutions.
- Research data: researchers don’t just work with subjects at their own institutions, or even just the US. How do we contract to guarantee compliance for EU citizen data, etc.
- Sponsored research: agreements might include terms governing patents, etc.
- Who can sign the agreement? Can the library sign it? Who would have to review?
- How does the IRB fit into the equation? Should the researcher pass on IRB agreement along with the data? Who assures compliance with the IRB?
- Missing BAA

- APTrust - climate is to minimize complications in every possible way. Drive assessment down to the depositor.
    - APTrust allows sub-accounts. UVA is “depositor” but can designate other related entities to deposit under that account.
- Internal systems - for any part of the system to take in sensitive data, the whole system must be able to. All of the agreements have to change.
    - Counterpoint from TACC: Not necessarily true. May be able to set things up so that not every component must be aligned.

Identify Elements of a DDP network for Private and Sensitive Data

Objective: Outline components of a service model, including a service proposition; key actors in design and governance; user interaction; technology and human resources needed for service delivery; associated costs; and metrics for evaluating performance.

Breakouts discussing and providing collective feedback about: (5 minutes thinking about it on your own, then 10 minutes at the table discussing and consolidating where possible, and then 15 minutes collectively)

- Technical Infrastructure
    - Where are there gaps in the existing DDP infrastructure or anticipated problems?
        - Over-reliance on cloud storage providers
        - Too much focus on ingest, not enough on restore functions
        - Workflows development
        - Proper storage/stewardship - still too many archives taking in materials that they can’t properly manage
            - Policies and procedures for intaking and handling information
        - Who is the ultimate security officer?
        - Lack of proper classification of data
        - Can the repository establish an overarching security model?
        - Encryption?
            - Who manages keys? Who makes sure they don’t get lost?
        - When there’s a legal battle over data ownership, to what extent will the network get involved?
        - Security analysis
        - Are there logically segmented locations for data with different protection needs?
        - Data portability -- how easily can you get on prem data to a cloud system, e.g.
        - Cost to maintain the infrastructure can be high and will be passed on to depositors.
        - Succession questions

        - Is data value defined?
        - Institutional memory - forgetting what's stored in the DDP
        - Can existing deposit tools be reused for the PII use cases?
            - Do we have two completely separate/parallel DDP systems?
        - Opportunity to do collaborative requirements-gathering BEFORE systems are built, to make it easy on depositors.
        - Complexity and lack of integration among existing infrastructures - REDUCE COMPLEXITY
        - Globus as common transfer tool
        - Staffing requirements - who controls/decide that? What if it requires radical staffing changes?
        - Who is in charge of auditing, ensuring compliance, etc?
        - Education for data depositors -- data ownership, roles and responsibilities, good assessment, good choices
            - Documentation for depositors
        - Legal agreements needed
        - No way for owners to indicate whether data is sensitive or not.
            - Metadata for describing restrictions
        - Versioning - OCFL as a possible approach? Could managers manage their own
        - Trustedci.org as a possible participant in these discussions
        - Insufficient scale
- Potential Service Models
    - What are the elements of good governance for this type of service? What does not work? *[Notes in this section need to be augmented from the sticky notes. -kp]*
        - Clear roles and functionality
        - Clear policies and regulations
        - Decision-making body made up of representative members
        - Board of stakeholders
        - Data-driven decisions
        - Standards-driven, willing to set standards for best practice
        - Use of new technology and no obsolete practices; use of existing tech but with improvements
        - Fosters collaboration
        - Good governance: provide owners with more control over management of data (incl versioning)
        - Grappling with succession of ownership responsibility
        - Shared vision
        - State-level agencies that govern mandates and compliance
        - Unaffiliated individuals who want to submit data documents
    - Who are the stakeholders and what are their roles?
        - Who owns the data? What is its use?
            - Data owners
                - Often a fight to determine the data owner (PI, university, etc.)
            - Considerations of cultural patrimony

            - In practice, the “owner” may be the institution
        - Depositors - but who is the depositor? Needs to be better defined.
        - Service providers
            - Chip’s vision: Facilitate connection to the “secrets service” - common front end to shared, distributed storage network
        - Repository managers
        - People in the data (person in a photograph, e.g.)
        - Future users of the content
        - Archives
        - Academic Deans
        - EVPs for health system and research for clinical and research data
        - Presidents for overall support/approval
        - University CIO
        - Developers
        - Transparency with communities ingesting data - communicating potential risks of, for example, reliance on commercial infrastructure
- Potential Cost Models
    - Where are the costs to consider?
        - More money in STEM than for cultural heritage collections.
        - Simplicity of structure may lead to lower costs.
        - Shared service facilitates economies of scale but also concentrates risk.
        - Many institutions moving to centralized model for services - take advantage of that.
        - HIPAA users have a lot of cash, so we should charge them.
        - Cost of assessment of materials
        - Nodes: software development, project management
        - Storage, servers, processing systems
        - Cloud service providers
        - Electricity bill
        - Security personnel, including physical security
        - Network transfer fees
        - Cost of key store/escrow
        - Auditing
        - Producing and maintaining documentation
        - Attorneys’ fees for agreements + money/time to negotiate contracts
        - Employee salaries
        - Data manager to select and prepare materials
        - Time/money for governance participants (sometimes stipends, often travel)
        - Maintenance of infrastructure
    - What are the potential models? (How do you want to pay?, etc.)
        - Public vs. private - cost can’t come from the library because then the campus thinks it’s free. Should be an institutional charge.
        - Tiered pricing based on, e.g., number of faculty.
        - On the questions of continuing payments vs. front-loaded (one-time) costs - need continuing payments to force continued commitment to the content

- Sustainability models - Jaime hasn't seen one that has ever worked. Difficult to propose and more difficult to implement. Easier in medical field where researchers are used to the "pay-for" model than in others.
- Hybrid Model: Join forces - institution will contribute funds, but individual researchers should share the burden.
    - This is the TACC model. UT System provides funding for infrastructure plus 5TB "free" storage for any researcher. Amazing how many researchers find that 5tb is exactly what they need.
- Funding agencies will pay for storage - but funding is time-limited.
    - Another issue - money from grants not a reliable source for preservation. Susan has experience with providing letter of support and writing in costs for preservation, but not in control of when the money comes to her from the project/PI.
- "Anything free will be abused."
    - Counterpoint - free stuff gets used. Lower the barrier to entry. From archivist point of view, there's a cart-horse problem. Everybody wants to know how many TB you need, but we don't know yet.
- Flat versus per TB cost. Libraries used to operating at smaller scale. Universities cut a $1 million check for gmail - no limit on storage.
- Consistency in cost is important. This is the reason many users go to a service like DuraCloud rather than using Amazon on their own - bc they can't handle/absorb the flux.
- Cost transparency
- Look at AWS as an example
- Service menu to adjust for cost for different needs

Closing and objectives for the remainder of the grant term

Notes