# Wrap Meeting Shared Notes

Friday, August 21, 2020

Meeting Agenda
project summary - Aug 2020
rough draft  final report - Aug 2020
Meeting notes (this document)

Background and prior meeting materials: Please feel free to share these materials with anyone who might be interested.

- Announcement of grant award, July 2019
- Meeting notes and presentations from a project partners meeting, Dec 6, 2019
- Mid-year update, Jan 2020

*This project was made possible in part by the Institute of Museum and Library Services grant # LG-34-19-0055.* https://www.imls.gov/grants/awarded/lg-34-19-0055-19

## Welcome and Introductions (11-11:30AM Central)

*Provide your name, institution, and one potential positive outcome of this project from your perspective?*

- Ashley Adair, Head of Digital Preservation and Stewardship, University of Texas at Austin
  *Better understanding of costs*
- Hesam Andalib, Graduate Research Assistant, Texas Digital Library
  *In line w/ TDL goals*
- David Bliss, Digital Processing Archivist, University of Texas at Austin
  *Concrete idea of whether the models would meet needs of sensitive collections*
- Bill Branan, Senior Engineering Lead, LYRASIS
  *Path forward for capturing data through Duracloud; value for institutions to be able to capture this data in light of few options*
- Sandeep Chandra, Executive Director for Sherlock, San Diego Supercomputing Center
  *Seeing how to help the Chronopolis project*
- Jaime Combariza, Director MARCC, Johns Hopkins University
  *Excited about the project*
- Bradley Daigle, Digital Initiatives Librarian and AP Trust Content and Strategic Expert, Academic Preservation Trust
  *New and fun ways to collaborate*
- Lea DeForest, Communications Manager, Texas Digital Library
  *Continued collaboration (ditto Bradley ^^)*

- Chianta Dorsey, University Archivist, University of Texas Southwestern Medical Center
  *Building DP infrastructure; excited to look at options coming available*
- Chip German, Program Director, APTrust, University of Virginia
  *Better common understanding of challenges in this space - want to be able to provide assurances out of substance, not out of "dreaming"*
- Heather Greer Klein, DSpace Product Manager, LYRASIS
  *Path forward for Duracloud users who need to store sensitive data*
- Ramona Holmes, Associate Director, University of North Texas Health Science Center
  *Standardized structure and foundational building*
- Christopher Jordan, Data Management Group Lead, Texas Advanced Computing Center
  *Would like to see DP environments take advantage of storage options for sensitive data for the whole data life cycle*
- Susan Kung, AILLA Archives Manager, University of Texas at Austin
  *Seeing what we come up with for sensitive cultural heritage (post-custodial)*
- Meg McAleer, Senior Archives Specialist, Library of Congress
  *Acquisition perspective - strong advocate to come up with capability to manage sensitive & restricted data*
- Isabel Meyer, DAMS Branch Manager, Office of the Chief Information Officer, Smithsonian
  *Raising awareness of options for storing secure / sensitive data*
- David Minor, Head of Research Data Curation, University of California, San Diego
  *Conversation moving to new spaces and new areas*
- Courtney Mumma (Project Lead), Deputy Director, Texas Digital Library
  *Starting conversations and excited to share*
- Francis Park, Historian, Joint History and Research Office, Joint Chiefs of Staff
  *Increasing # of holdings are digital and restricted, sensitive; looking to see what's available in the field*
- Pegah Parsi, Campus Privacy Officer, University of California, San Diego
  *Learning from project partners and share broader privacy perspective; initiatives around OA in scholarly data*
- Sibyl Schaefer (Project Lead), Chronopolis Program Manager, University of California, San Diego
  *Serve as foundation for standing up services for private & sensitive data*
- Lydia Tang, Special Collections Archivist, Michigan State University
  *Excited about the project & to learn more*

## Project objectives and work so far

*Review of the project's objectives and the work done to date. (Refer to project summary and draft report) Poll: Are you engaged with a DDP network?*

DDP definition: geographically dispersed and in at least three locations (refer to guide in slides)

Poll - depositor = 15%; service provider = 35%; not at all = 60%

## Objectives

- There are a number of DDP services that currently exist - today we have DuraCloud, Chronopolis and APTrust represented here, but there are also MetaArchive, LOCKSS, and others around the world.
- Yet, although distributed DDPservices have been offered in the United States for well over a decade, there is no distributed service offering for sensitive data. For this reason, the grantees propose that Personally Identifiable Information (PII) or Personal Health Information (PHI), as well as other sensitive data in the custody of libraries, academic health science centers, and archives is at an escalated risk of loss. Academic health science libraries, especially, face a growing backlog of digital PHI governed by HIPAA which requires preservation. Additionally, university-held cultural heritage collections are likely to have materials governed by FERPA requirements as well as valuable cultural heritage materials that contain PII such as social security numbers and other data deemed sensitive or private based on local and jurisdictional policies.
- Data specialists in academic health science centers and their colleagues in university special collections, archives and libraries share a common need for sensitive data preservation. The project team posits that the requirements for effective DDP of content containing sensitive and confidential information will be similar across all types of content. For example, the digital preservation services represented by the project partners and participants are content agnostic for anything BUT any content which is considered sensitive for any reason.

## Goals

- The project seeks to investigate the capacity and feasibility of a nationwide model for a DDP service that would close gaps in current preservation offerings for sensitive data for various types of institutions.
- The technology, infrastructure, and expertise needed to build a DDP service for sensitive data exist, but the connections, agreements and processes to put it all together to form a viable service are lacking. Since such an offering would be the first of its kind, it would be unwise to proceed to its creation before identifying the legal, technical, and financial requirements to build it effectively. While building this service is outside the scope of this one year planning grant, the two lead institutions are interested in using the grant deliverables to assess their capacity to meet the outlined requirements and to initiate discussions with possible network partners.

## Partnerships

- Both TDL and UCSD Library have established business models and years of experience building and providing distributed digital preservation services, as well as a history of collaborating with one another on these services. TDL and UCSD's relationship goes back almost 10 years ago, when we first started working together as nodes in the DPN network.
- UCSD manages the well known DDP service, Chronopolis. The Chronopolis network spans three sites across the United States and is one of the earliest established DDP services in the world.

- TDL, which is a consortium providing a number of services supporting digital libraries, has offered DDP storage services since 2015, and we became a replicating node for Chronopolis just a few years ago.
- Beyond the main project partners and IMLS, we've been grateful for the support and participation of more than a dozen institutions shown on this slide.
- These partners have, among other things, helped us collect use cases for private and sensitive data preservation and helped surface the technical, legal and service model needs and challenges.
- Partners listed on slide: TDL, UCSD, Chronopolis, IMLS, TACC, SDSC, Academic Preservation Trust (APTrust), the Smithsonian Institute, Northeastern University, the University of North Texas Health Sciences Center, the University of Texas Southwestern Medical Center, the Dell Medical School at the University of Texas at Austin, DuraCloud/LYRASIS, and the Maryland Advanced Research Computing Center (MARCC) at Johns Hopkins University

## The HIPAA bar

- The project team hypothesized that the bar set by HIPAA requirements is sufficiently high to protect many other kinds of nonregulated sensitive data.
- So based on that bar, what we are really looking at is what it takes to provide a HIPAA-compliant DDP network that will accommodate all the other kinds of PII I've mentioned. Note that we are deliberately excluding classified and top secret data.

## Key resources

- Both TDL and UCSD Library maintain close working relationships with data centers affiliated with our home institutions that could provide key resources for a DDP service for sensitive data.
- The Texas Advanced Computing Center (TACC) at UT Austin has partnered with TDL on several projects over the years and already provides storage resources for the TDL's Chronopolis DDP storage node. For HIPAA (Health Insurance Portability and Accountability Act) and FERPA (Family Educational Rights and Privacy Act) secure and compliant storage, TACC has undergone a rigorous audit and costly compliance upgrades to accommodate local customers.
- Similarly, the San Diego Supercomputer Center (SDSC) provides HIPAA compliant storage to UCSD faculty and researchers. Both computing centers have shared with us their expertise in providing sensitive data solutions, including details of their services and related costs.

## Work so far

- We started last summer, with hiring our GRA Hesam Andalib. Hesam hit the ground running and gathered a number of use cases and outlined the agreements between all the parties in the Chronopolis network.

- We then met with all of our partners in Austin last December, back when traveling and meeting in-person was still feasible. We gathered some very valuable information from our partners during that meeting.
    - The grant partners convened in Austin in the end of 2019 to discuss the need for private and sensitive data DDP services, as well as the legal and technological challenges. We agreed that many institutions need such a solution since they are holding or planning to accept, for example, student records, email, health data, human rights archives including first person accounts of trauma, and commercially restricted data. Personally Identifiable Information (PII), Personal Health Information (PHI) and other sensitive data that has significant preservation value in libraries, archives, and academic health science centers is at an escalated risk of loss.
    - We reviewed our case studies and confirmed that it is common practice for archives to refuse any data that contains PHI or PII, regardless of its historical or evidential value because they don't have the means to steward it or the resources to manage it. In some cases, the authority to manage materials properly is unclear and even if accessioned, there is nowhere to put them.
    - And in other cases, we know that institutions haven't done the level of assessment or appraisal necessary to quantify the problem sufficiently at all, and there are often issues clarifying the owner of the materials, etc.
        - This is very common for archives w/ data that aren't yet processed.
- Since then, we've worked to determine gaps in the current agreements and processes (both internal to DDP networks and any user barriers) that should be addressed to satisfy the requirements of private and sensitive data.
- We've also investigated service and cost models and will be disseminating our findings at the end of the grant term.
- Final report is drafted - feedback is welcome.

## Q&A

Jaime Combariza:

- Agree w/ what was represented and see challenges w/ legal requirements / different requirements at different institutions.
- Cost: Researchers at JHU have to pay for storage. They need to know in advance and chances are their grants are over by the time they get to the preservation stage.
    - Response:
        - Courtney: Challenge of building in DP as part of infrastructure is a challenge across the board. Cost will vary between institutions so it's hard to standardize. Legal complications are absolutely true; difficult to download requirements and develop definitive templates. Legal expertise is needed at every point of the process, but resourcing legal and contract resources are scant among libraries and archives.
        - Kristi: Legal hurdles and need for resources to clear the legal hurdles still need to be figured out.
        - Sibyl: Cost will vary by provider, e.g. TDL has a different way to price out HIPAA-compliant storage than other consortia. Worthwhile goal to collaborate on standardizing prices. Angle to approach: develop guidelines for the best way to care for the data.

Susan Kung:

- Presented a white paper to express the need for researchers to obtain funding for archiving research; paper was well received among major funders (NSF, NEH, etc).
- Important to include at the time the grant is written.
- Major funders will respond well to costing models; program officers needed the explanation to why preservation/archiving is needed - they just needed to know.

## Service options

*Review of service options (refer to project summary) including technical requirements and costs.*
*Poll: Select all the services that your institution might join and/or participate in.*

Based on our investigations to date, and on the goals as presented, we came up with a few service model options which we briefly described in the project summary shared late last week.

- Each model assumes that the service provider is already a DDP service provider, and runs a service that is in alignment with digital preservation good practice.

### Single Node Offering (HIPAA)

- Assuming that there are depositors currently responsible for data covered by HIPAA and which is ready for preservation storage, both UCSD and TDL could independently offer a fully HIPAA-compliant single storage node solution to their respective depositors by leveraging their existing partnerships with San Diego Supercomputer Center (SDSC) and Texas Advanced Computing Center (TACC), both data centers affiliated with our home institutions that could provide key resources for a DDP service for sensitive data.
- TACC, for instance, already provides storage resources for the TDL's Chronopolis DDP storage node. For HIPAA and FERPA (Family Educational Rights and Privacy Act) secure and compliant storage, TACC has undergone a rigorous audit and costly compliance upgrades to accommodate local customers.
- SDSC provides HIPAA compliant storage to its faculty and researchers at UCSD.
- Both computing centers have shared with us their expertise in providing sensitive data solutions, including details of their services and related costs.
- In this single node service model, in order to achieve true DDP status, the depositors would leverage local storage for one or, preferably, 2 geographically distributed preservation copies of their content.
- The Texas Digital Library's Digital Preservation Service, for example, already recommends that any institution using the storage provided by TDL have at least one local copy which replicates the content they've ingested into the TDL systems. This is for two reasons:
    - first, it's so that institutions using Chronopolis via TDL don't have to pay fees to access their stored content
    - second, it's so that institutions using the Amazon options TDL provides have a second or third node available as well as the closer access without egress costs.
- If TDL were to offer a HIPAA/FERPA TACC storage option in its Digital Preservation Services, its members could use it via DuraCloud at TDL as the third secure location in addition to their institutions' copies.

    - Since most of the members with medical libraries and health science centers have information technology departments accustomed to managing PHI in their regular course of business, their archives and libraries can presumably leverage those functions for two locally distributed digital preservation copies by provisioning areas for their materials over which they have control.
    - Relationships between archives or libraries and their associated information technology units can be complicated, but TDL already offers consulting and strategy support to Digital Preservation Services members to assist in technological support and storage provisioning in addition to advice for processing and preparing PHI content for preservation.
- TDL would offer such a service under mostly the same governance and pricing structure as its current services, revising current Digital Preservation Services SLAs to allow for private and sensitive data and confirming that Business Associate Agreements (BAA)s are in place appropriately. This would likely be the least costly option, as the service charges would only be for one node of the DDP network.

Question from Bill Brannan: Courtney, could you talk about what you mean by "depositors are responsible". What do you see depositors needing to do beyond keeping copies on their own?

- Response:
    - As part of TDL's DPS services, we would want depositors to be responsible as part of the contract. The next step of the process would be to ask them to agree to manage the copies. HSC and med libraries have access to services that support secure storage, but TDL would not have dominion over copies.
    - Outline roles & responsibilities before content moves into HIPAA-compliant DDP storage.
    - Coordination with the institution and service provider to reconcile # of copies and checksums are checked on a regular basis, etc.

Question from Jaime Combariza: What would be the process to access, if needed, these additional copies of data. One "contact point" or someone at SDSC/TACC?

- Response:
    - It's easier for institutions to keep one local copy so that it's easier to access the copy and avoid costs associated with egress.
    - If the copy is lost at the local institution?
        - For TDL, one point of contact and we would retrieve from AWS storage. For Chronopolis, UCSD Library would manage through Sherlock; point of contact would be administrators to retrieve from storage.

## Full DDP Offering (HIPAA)

- A fully HIPAA-compliant, 3 node DDP network service model would be the most complex, and likely most costly of the options we have discussed. Suffice it to say this service model would require a lot more preparation.
    - Including, for instance, that the two lead project partners have only identified two potential nodes. We could, however, make moves to incrementally connect these two

    nodes as a start to a three-node, HIPAA-compliant DDP network and as an entry point to initiate discussions with possible technical and community network partners.
- Both UCSD and TDL have access to their own data centers with HIPAA storage.
    - If the two partners could find a third node with similar access to a data center with HIPAA storage, there would be enough parties to participate in a Chronopolis service option to accommodate private and sensitive data.
    - Chronopolis has established partnership documentation and processes, including reciprocal agreements, so the structure to accommodate a new partner exists.
- The nodes at SDSC and TACC have different associated costs, and a new node might also have a cost difference, so there would be some work to align costs during negotiations.
    - The service would be governed and maintained under roughly the same structure as the current partnership between Chronopolis and TDL and their associated partners.
    - Legal binds would have to be amended and corrected to accommodate private and sensitive data, PHI and student data.
    - All parties would likely need to consult with their respective privacy officers and legal departments to ensure that all the requirements to accommodate essential PHI and PII were fulfilled according to HIPAA guidelines.

Questions re: Full DDP Offering (HIPAA):

- Bill Brannan: Anything from a technical perspective that would be different from the Chronopolis offering?
    - Response:
        - More to come on encryption.
        - Most of the issues come from Legal questions.
        - Most technological infrastructure is already in place, but because of that, the tech infrastructure complicates the legal agreements.
- Chianta Dorsey: a 3 node network = items are stored in 3 geographical locations?
    - Response: (Courtney) Yes. Difference in this full DDP offering is that the distribution would be handled by the service itself.
- Chip German: Perspective on liability in the development of the models. People are still depositing at their own risk. Do you recommend a different approach re: insurance for damaged or lost copies. Example: DNA history - record preserved but “lost”; generations of encryption may render copies unreadable.
    - Response:
        - Liability currently does not live w/ service providers but is the depositors’ responsibility.
        - More research to be done around liability and insurance
        - Real liability would be unauthorized access to the data and that would be harder to insure
        - Liability is a gap to be researched; how we do it is TBD. Certainly the cost of the offering would be increased.
        - With no PHI, liability is placed on the depositor but the risk is low. If you say you’ll take PHI, the risk increases, and vigilance must increase accordingly.

- Bill Brannan: Considering the need for a 3rd node to fill out the network, what limitations do you see in using commercial (cloud?) providers as part of this role? Using commercial cloud as a node.
    - Response:
        - Courtney Mumma: Top tier best DP network will be provided from within the higher ed/cultural heritage community. What differentiates us from commercial providers & government is a good practice issue.
        - Isabel Meyer: It will depend on the commercial cloud provider: Cost involved in frequency of access, # of copies, level of protection. At the Smithsonian, they are still using internal storage because commercial storage is too costly
        - Chip German: APTrust would architect a solution from the perspective of the community in mind. Considering the time out - trying to train yourself to think at the 100 year level and not the five year level. Interested to learn about thoughts around generations of encryption. Only assumption we need to make is that we need to imagine doing something different than what we're doing right now (still on the five year thinking).

## HIPAA-like DDP Offering

- Another option is creating a HIPAA-like DDP network, meeting all the requirements but without the costly audit and certification at every part of the data flow.
- In the HIPAA-compliance universe, this is sometimes known as being HIPAA-eligible.
    - A service that is HIPAA eligible is one that is capable of being configured in a way that could meet HIPAA compliance requirements. HIPAA itself doesn't really provide technical details, it just indicates that a service provider implement technical security measures to guard against unauthorized access to electronic protected health information that is being transmitted over an electronic communications network.
    - So it's really the extensive documentation and auditing process that take a service over the bar to achieve HIPAA compliance designation.
- Because we've found that so many of our partners aren't actually accepting real HIPAA or FERPA data into their collections, we would need to redefine HIPAA's 'covered entity' (the depositor with private and sensitive data) and 'business associate' (who handles the data) for our own purposes for those non-HIPAA certified elements of the data flow.
- Really, this option might actually be HIPAA compliant, just without the costly auditing and certification of every cog and gear in the process. This would be less costly for the depositors than the fully HIPAA compliant service model for that reason.

## Fun with Encryption

- For each of these service options, it's up to the service provider whether they wish to encrypt data or not. HIPAA states that PHI in transit and at rest need to be secure, but doesn't explicitly regulate how.
- In digital preservation circles, encryption is not considered in alignment with good practice, though it has been very difficult to identify sources explicitly discussing why that is. My own understanding is that it disrupts transparency, amplifies the risk of losing access to content in

the future when keys go missing, and impedes the ability to monitor content over time to do preservation planning and perform migrations.

- Whatever the reasons, even if a DDP service provider chooses not to encrypt or to ask their depositor to encrypt and be solely responsible for their own keys, the service provider is still liable to protect the data if it wishes to be HIPAA compliant.
    - If the provider decides not to encrypt at all, they must still provide robust justification about how they provide the same level of security without encryption AND why it was not applied.
- So, for DDP providers, at least based on the investigation so far, we'd need much better evidence about why encryption doesn't align with digital preservation good practice.

Questions re: Encryption:

- Francis Park: Has there been any discussion of public key infrastructure in encryption? For emails, that is
    - Response:
        -

### Considerations

Some things to consider when thinking about how to move ahead with service provision:

- Is there actually HIPAA-covered data which is at a state in the digital preservation lifecycle workflow that it is ready to move into digital preservation storage?
    - And if so, is it enough to justify a service. And if it is enough to justify a service, which service model is the most reasonable for now?
- Does the potential service provider have resources available to it to engage legal advisors for changes to contracts and agreements; compliance officers and funding; security expertise; developers; technical documentation writers?
- And, finally, what resources are available to the depositors? So many of the institutions we spoke with throughout the course of the past year aren't even engaged with DDP networks for their typical digital preservation needs.
    - Do the depositors have digital preservation expertise and if not, can they manage the resources needed to get good training?
    - Once they're trained up and have policies and procedures in place, do they have the funding to do the extra layers of processing necessary for digital preservation and/or for sensitive data?
    - How will they manage acquiring access to the skills necessary to mitigate risk involved with handling HIPAA and FERPA content?

## Legal binds and other considerations

*Review criteria & legal binds between parties and other service model requirements for each service option.*

Now that we've introduced the three service options, let's get into some more detail about the criteria and legal attachments between the involved parties and other service model considerations.

## About HIPAA

- Today’s meeting, and the report in general, is not intended to be a master class in HIPAA. This illustrates the very minimum, basic configuration and requirements of a HIPAA-compliant DDP network. It includes the essential roles of ‘business associate’ and ‘covered entity’, which I briefly mentioned before but which I’ll now define more fully.
    - The covered entities are health plans, health care clearinghouses, and health care providers who electronically transmit any health information in connection with transactions for which the Dept of Health and Human Services has adopted standards.
    - A “business associate” is a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of, or provides services to, a covered entity.
    - Understanding these two roles is essential to understanding HIPAA-compliance, and attributing those roles to parts in a DDP network has proven to be core to planning for a service.
- We won’t address too much about FERPA today - that’s because most of the FERPA requirements are compensated by documentation and contract language.

Questions about HIPAA:

Pegah:
At the risk of causing people to throw things at my head, I’m wondering if GDPR needs have also been considered since they are different than HIPAA?

Courtney:
We have not gone to GDPR at all

Sibyl:
DDP network does not know what kind of data is coming into the network. If somebody wants some extraction from DDP data they need to go to depositor.

## Node Models

### 1. *Single Node Offering (HIPAA)*

The simplest of our suggestions is the single node offering. Both TDL and UCSD have everything they currently need to move ahead relatively immediately with such a service.

### 2. Full DDP Offering (HIPAA)

The second suggestion is the fully compliant 3 node DDP network. This is basically the application of HIPAA compliance requirements across the existing Chronopolis network, for example. But we need a third node.

*See slide 28:* There are many connections which require legal agreements of various kinds (contracts, SLAs, BAAs, MOUs, and so on), and that creates a barrier to the formation of a DDP network for sensitive data that strictly follows HIPAA requirements.

*See slide 29:* Over the course of our work this year, we have been considering how to overlay our existing DDP network infrastructure data flow, which you see here, with the needs of the HIPAA one and identify the requirements to accommodate a subset of sensitive data without disrupting the regular service or undermining compliance to HIPAA as the gold standard. Currently, SDSC and TACC are the only compliant nodes we've identified, so we're already at a bit of a disadvantage in terms of being able to form a DDP network with a minimum of 3 distributed nodes. We also need more documentation and considerable changes to current legal agreements at every data access point in the network workflow.

### a. For either HIPAA offering

The extensive technical and legal connections required to enable our current services and which will be needed to move forward with DDP or single node options are not insurmountable.

i. They are, however, added complexity which only becomes more so when you add audit and certification requirements for partners and new nodes across different jurisdictions and boundaries.

ii. For either of the HIPAA-compliant offerings, we would need to begin by making sure that our agreements no longer exclude private and sensitive data, HIPAA or FERPA data, but do exclude secret, top secret and confidential data as defined by the US Government Classification System. For FERPA specifically, we'd also need a statement of FERPA compliance.

iii. In addition to updating each agreement between all involved entities, we'd need BAAs for any business associate in the service model which will receive access to PHI at any point in the workflow from a HIPAA depositor (aka 'covered entity'). Right now, in our current DDP models, I can access our systems and our depositor's content from my home at certain stages in the data flow, which most certainly not compliant with any regulations, for instance. We'd need to restrict access to PHI for any individual or service entity at a location not covered by a BAA.

iv. It would also be unwise to move ahead with any service provision for HIPAA and FERPA data without regular access to legal expertise to review the agreements and monitor compliance. We'd also need to engage with technical writers to provide justification for not using encryption should we decide to abstain from it in alignment with good digital preservation practice. Further, we need to better explain why it's not good practice using evidence from experts in the field.

## 3. HIPAA-like DDP Offering

Because of the costs associated with HIPAA certification and the variation in the resulting costs of storage, service partners could consider providing a less expensive option that mimics HIPAA storage requirements for other kinds of private and sensitive data which is not governed by that legislation. This would include PII in manuscript collections, archives and libraries discussed at length in the use cases discussions we touched on earlier and which are outlined in the project summary and draft report.

a. Such an offering could still leverage HIPAA storage like that at SDSC and TACC; however, the various partners and systems engaged when data moves in and out of the system would not have to undergo HIPAA audit and certification. DDP services like Chronopolis already provide a high level of security during ingest and replication.
b. Because we've found that so many of our partners aren't actually accepting real HIPAA or FERPA data into their collections, we would redefine HIPAA's 'covered entity' and 'business associate' for our own purposes for those non-HIPAA certified elements of the data flow.

### HIPAA-lite/ish/like checklist

- ❑ Do not exclude private and sensitive data in agreements
- ❑ Exclude HIPAA data
- ❑ Exclude Secret, Top Secret and Confidential under the United States Government Classification System in agreements
- ❑ Legal agreement between Service Provider and Storage Provider
- ❑ Legal agreement between Depositor and Service Provider
- ❑ Legal agreement between Service Enabler(s) and Service Provider
- ❑ Provide rigorous information security documentation for every stage of the deposit lifecycle

Our checklist of requirements to consider adding to your DDP model for this service model are very similar to those in the HIPAA-compliance list. Really, what we'd be doing is that HIPAA-eligible level of treatment across the board providing better documentation and compensating for the private and sensitive data content in our agreements, but saving ourselves the audit and some of the more restrictive access policies.

## Discuss the 3 service options

*Group discussion of the 3 service options presented.*

*Breakout rooms to answer these questions: Report back (20 minutes)*

- *Discuss barriers and advantages of any one of the service options for your own institution. (10 minutes)*
- *What capacity building activities are needed at your institution to be ready for a DDP for sensitive data? (10 minutes)*

## Group 1

*Discuss barriers and advantages of any one of the service options for your own institution. (10 minutes)*

- UVA - from the depositor perspective; Cost is a barrier- especially matching with the idea that measurement by tb and not mb (large scale is essential to be worth the effort); academic records' storage and management is still an unknown; new content refresh
    - As provider - the resources for review and preparation are significant and would be passed on to the provider
    - HIPAA-lite is the most feasible from the APTrust provider
- Chris -TACC - obstacle if you say HIPAA-ish/etc, then the immediate question is 'how is it different?"; TACC has applied the HIPAA as the high bar as our project has done; from a legal and procedural this makes things easier as they are already compliant
- Chianta - from the UTSW-Med - HIPAA-ish wouldn't work at all; full compliance is a necessary, which is why Chronopolis hasn't been a consideration; HIPAA data is in the physical records and will be in the digital records for sure; unprocessed BD collections might have PHI/PII; Barriers - risk averse due to the nature of the institution; info resources component deal with all technology across campus including the hospitals - they want to give approval for any new technology, software, on the campus;
    - Single Node feasible for UTSW - already successful with having two locally; 3 node options more of a hurdle because of all the certifications
    - While considering dark archive - issues with AWS are the unknowns about how they manage fixity and access; TDL offering would be more trusted in those areas wrt transparency
- David - either HIPAA or Not for the purposes of UCSD; for single node replication - the biggest challenge would be the cross-node audit/alignment; how do the local copies
    - What kind of dashboard would we need for the management of data across all of the nodes in the single node model.
    - HIPAA ish - liability issues, risk issues; are we prepared to pay the cost of the intensive audit requirements

*What capacity building activities are needed at your institution to be ready for a DDP for sensitive data? (10 minutes)*

- Knowing the risk and liability is essential, even moreso than 'guarantees' of safety and security
- Ongoing audit requirements for HIPAA - security controls and regular audits for compliance (TACC); keeping docs up to date
- TACC, etc - broadly identification and management of all data to assess requirements for what controls apply - or event o have the 'buckets' for data types; lacking in most cases; institutional capacity to manage and sort content

- UTSW-Med - convincing IR about WHY such service is needed - business needs justification

## Group 2

*Discuss barriers and advantages of any one of the service options for your own institution. (10 minutes)*

- Compromise between cost and all other concerns
- To what degree will cost sharing be an issue? The current partnership agreement is a reciprocal storage agreement. How does HIPPA compliant storage challenge this arrangement?
- Many more people have adjusted to cloud storage for sensitive date due to the pandemic.
- For the most robust option, is there enough readiness and enough demand to make it worth the initial large outlay of cost and effort for the amount of use?
- There is a need for HIPPA compliant data storage, but are they looking for DDP?
- HIPPA-like option had the most interest. Why? What use cases?
    - Having a HIPPA-like environment would make access to preservation storage and continuity of local storage --the  safeguards are in place - would help.
- LYRASIS clients - only heard from institutions that want to throw everything in there and not have to worry about assessment. Thinking about PII primarily.
- Smithsonian: I don't deal with HIPAA data in our repository at all.

*What capacity building activities are needed at your institution to be ready for a DDP for sensitive data? (10 minutes)*

- One reason for the question is, in past projects institutions were not ready financially or institutionally for the service. Problem magnified here when some institutions are not collecting this data at all.
- Awareness-building of those who generate the data of their responsibility.
- Hard to do alone without partners, hard to get started or ready.
- Collective assessment education, collection assessment for PII. There may be organizations who don't know they already have stewardship of data that meets this criteria.
- Educating funders about this is another activity. NIH, IMLS, and others to ask them to build it into project proposals.
- Don't have to choose one. Could start with one, and build capacity and infrastructure for PII storage and move to something more robust.
- Options - HiPAA like was attractive for those who don't have HIPAA or FERPA, easier to convince people that would be what they need to tribal or IRB requirements. Especially for non-covered entities.

## Group 3

*Discuss barriers and advantages of any one of the service options for your own institution. (10 minutes)*

- Trusting non-commercial products. Hospital mindset, want something established that peers are using.
- Difficult to get buy-in from institutions that this is necessary, that we are managing materials that we don't own. If we aren't the owners then less interested in footing the bill if we don't have copyright or rights to the material
- Concerns with unauthorized access and making guarantees about it when they aren't the administrators of the storage
- Distinction between covered entities vs non-covered entities - non-covered find HIPAA-like more attractive, covered need full HIPAA guarantees
- GDPR - feeling like that's a European law, we don't need to worry about it in Texas
- Margaret McAleer, Ms div of LOC - they have an internal DDP w/ 3 distributed nodes, but they don't actually control the servers in their DDP, they control the content in and out, but they cannot restrict access to the service. EX: depositor wants them to guarantee there can be no unauthorised access, but they cannot make that guarantee. They have not gotten the material yet. Library IT staff manage the servers, enterprise level. Trust issue on the part of the depositors.
    - Sibyl: Would a certification of trust help?
    - Margaret: Yes
- Susan & Margaret: It's difficult to get the rest of the institution onboard and make the case that this level of service is necessary. Especially wrt materials on deposit and that the institution does not own for IP for.

*What capacity building activities are needed at your institution to be ready for a DDP for sensitive data? (10 minutes)*

- Have management systems in place, would need to have migration factors, standardized structure for moving data.
- Ramona: Migration plan to move data into such a DPP. The existing products are "kind of a mess" and there is no way to migrate from one to another.
- Migration paths are always difficult
- Margaret: we have a well-established dpp workflow. The difficulty will be putting the policies and agreements in place.
- Sybil: UCSD doesn't have any sensitive data that is processed - where is the data? What are the processes for putting it into a system? Liability? Clarity around roles of depositors vs service. It all comes down to the legal agreements.

# Next Steps discussion

*Discuss the next steps including further dissemination and implementation plans.*

We've left ample space in our draft final report to include information we've gathered from you at this meeting. We also have to do some more synthesis of research we've already done.

We have extended the term of the grant for another year primarily because we did not end up using funding which had been allocated for dissemination and discussion in person at various annual conferences. We are looking for ways to continue sharing our findings and getting feedback virtually, as well as looking at some professional development skills that will contribute to our findings and the work going forward.

## Users | Ownership | Cost

Even beyond the complex legal underpinnings of current DDP services and those required of HIPAA-compliance, a **lack of user readiness**, **challenging questions about ownership & liability**, and the **high cost of HIPAA certification** undermine the feasibility of a service offering that accommodates private and sensitive data that is not PHI.

### Lack of User Readiness

- During the discussions of use cases, our partners and the project team struggled with defining what qualified as sensitive and private content, even with access to our own institution's data classification standards.
- Additionally, there was little clarity about the level of protection data required across different individual units, between institutions, from state to state and across national boundaries. Determining control and ownership of content is complicated, and most of the institutions with whom the project team engaged throughout the term of the project lack the capacity and/or resources to properly determine the extent of private and sensitive data at risk in their possession.

### Questions About Ownership & Liability

- Further, some of the partners do not acquire or accept transfer of any content that may contain such data since they know they do not have the means to protect it or to provide properly mediated access to it. For them, it's simply not worth the risk.
- The overall perspective of the project team is that most institutions are not ready to acquire, process and preserve private and sensitive data. DDP storage is the final stage in a complex process from accessioning, processing and preparations for ingest through a digital preservation workflow. Without maximal effort towards the beginning stages of the process, the need for DDP storage seems far off for all except a very small group of preservation professionals.

### Cost

- With complexity and the challenges of more rigorous examination of our expertise, practices, interconnections and documentation, comes added costs. Those costs differ depending on the storage facility, for instance we found that the pricing differs between TACC's and UCSD's. Nevertheless, the costs are still more expensive than 'regular' storage, which means that institutions with sensitive data requiring the top tier DDP storage option would likely only want to store the content that is at-risk.
- One glaring issue is that this practice of isolating PII data from the context of its collection is incompatible with the way that most archives manage their aggregations. At least intellectually, archives are described using a simple hierarchy and not down to the item level. Unless archives are resourced so that they are capable of analyzing sometimes massive collections of items to determine the location of the individual records containing private and sensitive data, they will have no choice but to select from a higher level of description containing items likely to contain such data and place it in DDP sensitive data storage at a higher cost.
- In regular DDP networks, we also frequently see depositors depositing unprocessed material that they intend to process at a later date. If there is a chance that those collections contain PII or PHI, there would be no choice but to choose the more expensive, safer DDP storage option.
- Finally, we also face the problem of an ever evolving legal environment. Data is preserved at a single point in time in that evolution. There are implications for DDPs when data was received under one expectation of privacy that changes as the legal environment does over time.

## Next Steps

Over the past year, what the project team has learned will inform any service providers wishing to move ahead with developing DDP services for private and sensitive data.

- For TDL and UCSD, which service offering is the most feasible?
  - Start one node offering
  - Grow internal expertise HIPAA
  - Work towards 2 node DDP as HIPAA-like ('HIPAA-eligible')
  - Secure 3rd node
  - Become fully HIPAA compliant across the entire infrastructure
- As a community
  - How do we prepare institutions to accept/take custody of private and sensitive data, including HIPAA and FERPA content?
  - Is there another grant needed to move the needle on this process? If so, who would be involved and what are the goals?
- Continue dissemination & discussion

For TDL and UCSD, we need to make decisions about which service offering is the most feasible to begin? We could start with unique, one node offerings while we expand our knowledge base and

HIPAA expertise. Over time, we could work towards a 2 node HIPAA-eligible network as we look to secure a 3rd node…. Any takers??

Eventually, we could become fully HIPAA compliant across the entire infrastructure.

As a community of digital preservation practitioners, we also need to take a big step back and consider how to properly prepare institutions to acquire private and sensitive data. Readiness is still a massive problem even in regular digital preservation. Do we need another grant to move the needle? If so, who would be involved and what would the goals of that effort be?

### Your ideas / questions

- APTrust is happy to stay part of the conversation.
- Bill Brannan - what professional education opportunities are we considering?
    - Sibyl is looking at a HIPAA certification class (25 hrs + test)
- Project Leads:
    - Are looking for opportunities to share / disseminate findings and we have funds to spend from our travel budget on conference registrations.
    - Are looking for that third node and future project partners on the next grant
        - Susan Kung: Have we looked outside of the US?
            - Could satisfy GDPR issue?
            - Would only apply to the HIPAA-like service model
            - Meeting the highest bar with different issues
    - Two levels of data classification systems - UC System and UT Systems
        - Develop a crosswalk?
        - Share data classification documents

*This project was made possible in part by the Institute of Museum and Library Services grant # LG-34-19-0055. https://www.imls.gov/grants/awarded/lg-34-19-0055-19*