# Preserving Sensitive Data in Distributed Digital Storage Networks

Texas Digital Library (TDL), in collaboration with the University of California, San Diego Library (UCSD)

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

The Texas Digital Library (TDL) and the UC San Diego (UCSD) Library initiated a project in 2019, funded by an IMLS grant, to explore the feasibility of a nationwide model for the first Distributed Digital Preservation (DDP)[1] solution for sensitive and protected data. What the project team has learned will inform any service providers wishing to move ahead with developing DDP services for private and sensitive data.

Although distributed digital preservation (DDP)[2] services have been offered in the United States for over a decade, no distributed service offering for sensitive data currently exists. There are several reasons such a service has not been established, not the least of which are liability concerns and other legal ramifications which make it difficult and complex to establish. As a result, Personally Identifiable Information (PII) or Personal Health Information (PHI), as well as other sensitive data in the custody of libraries, health science centers, and archives are at an escalated risk of loss. Academic health science libraries, especially, face a growing backlog of digital PHI governed by HIPAA. 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. It's also regular practice for archives to refuse any data that contains PHI or PII, regardless of its historical or evidential value simply because they don't have the means to steward it.

The project team hypothesized that the bar set by HIPAA (Health Insurance Portability and Accountability Act)[3] requirements is sufficiently high to protect many other kinds of nonregulated sensitive data. Based on that bar, the Preserving Sensitive Data in Distributed Digital Storage Networks project team has examined the feasibility and requirements for a nationwide DDP service that would close gaps in current preservation offerings for sensitive data by considering what it takes to provide a HIPAA-compliant DDP network which will accommodate most other kinds of PII, but exclude classified confidential, secret and top secret data[4].

With the support of our project partners, the team gathered research and data needed to model a nationwide distributed digital preservation service for private and sensitive content. This report will discuss the foundations needed for the establishment of such a DDP service, including the technical requirements for data transfer and cost modeling scenarios. Based on the findings of the grant, TDL and UCSD intend to incrementally enhance their own current DDP offerings to include services for sensitive data. We hope that our examples, along with the information and recommendations provided here, will pave the way for other DDP services to do so as well.

---

[1] Distributed Digital Preservation (DDP) Definition from MetaArchive Cooperative. A Guide to Distributed Digital Preservation.Atlanta: University of North Texas Libraries: 2010. digital.library.unt.edu/ark:/67531/metadc12850/. Accessed August 12, 2020.

[2] Distributed Digital Preservation (DDP) Definition from MetaArchive Cooperative. A Guide to Distributed Digital Preservation.Atlanta: University of North Texas Libraries: 2010. digital.library.unt.edu/ark:/67531/metadc12850/. Accessed March 14, 2019

[3] The Health Insurance Portability and Accountability Act (HIPAA). Washington, D.C.: U.S. Dept. of Labor, Employee Benefits Security Administration, 2004.

[4] United States Government Classification System. Issued by President Barack Obama in 2009, Executive Order 13526 replaced earlier executive orders on the topic and modified the regulations codified to 32 C.F.R. 2001. "Executive Order 13526 - Classified National Security Information". Information Security Oversight Office of The National Archives. Retrieved January 5, 2010.

## Texas Digital Library and Chronopolis Distributed Digital Preservation

Both TDL and UCSD Library have established business models and years of experience building and providing DDP services, as well as a history of collaborating with one another on these services.

The Texas Digital Library (TDL), administratively based at the University of Texas at Austin (UT), is a consortium of Texas higher education institutions that builds capacity for preserving, managing, and providing access to unique digital collections of enduring value. The mission of the TDL is to advance and advocate the role of digital libraries and digital scholarly communication technologies that support the research and teaching missions of institutions of higher education in Texas and to promote cooperation, communication, and resource sharing among its members.[5]

Since 2015, the TDL has also offered access to DDP storage systems.[6] Early iterations of its digital preservation services allowed members to store and manage multiple copies of data in Amazon Web Services storage locations and/or at the Texas Advanced Computing Center (TACC). In 2012, the TDL joined the Digital Preservation Network (DPN) and worked in partnership with UT Austin and TACC to build and launch in 2016 one of four production nodes in that network, which ceased operations in 2019. In 2017, the TDL joined the Chronopolis DDP network headquartered at the University of California San Diego Library, providing access to Chronopolis services to its members (via TDL's DuraCloud implementation) and serving as a replicating node for the network using storage at TACC.[7]

The University of California, San Diego Library manages the internationally-recognized 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, having been in operation for well over a decade. The UCSD Library partners with the University of Maryland Institute for Advanced Computing Studies (UMIACS) and the TDL to maintain geographically distinct data centers. Chronopolis offers preservation storage through the DuraCloud and TDL services. It was certified as a Trusted Digital Repository by the Center for Research Libraries in 2012.[8]

Both project partners maintain close working relationships with organizations affiliated with their home institutions that could provide key resources for a DDP service for PII and PHI. Both TACC, located at UT Austin, and the San Diego Supercomputer Center (SDSC), affiliated with

---

[5] "Texas Digital Library Bylaws." https://www.tdl.org/wp-content/uploads/2018/05/TDLBylaws_201805.pdf. TDL.org. Accessed March 16, 2019.

[6] "Announcing DuraCloud @TDL for digital preservation." TDL.org. November 12, 2014. https://www.tdl.org/2014/11/announcing-duracloud-tdl-digital-preservation/

[7] "Texas Digital Library Joins Chronopolis Digital Preservation Network." May 11, 2017. https://www.tdl.org/2017/05/texas-digital-library-joins-chronopolis-digital-preservation-network/

[8] International Organization for Standardization. *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories.* ISO 16363:2012 (CCSDS 652.0-R-1). Accessed March 14, 2019. https://www.iso.org/standard/56510.html. ISO 16363 is the highest level of digital preservation certification available.

UCSD, offer protected data storage. These individual storage locations fall short of providing the recommended three geographically distributed copies, but could serve as essential components as the project team works to model a best-practice DDP network for PII.

In carrying out their missions and supporting their current members, both the TDL and Chronopolis have observed that data containing Personally Identifiable Information (PII) or Protected Health Information (PHI), as well as other sensitive data managed by libraries, academic health science centers, and archives, are at an escalated risk of loss. Consultations with TDL member university libraries and archives reveal that more than half of all 22 member institutions have sensitive data content which requires digital preservation actions, and a recent survey of University of California libraries found that over half of the UC libraries manage HIPAA data. While some digital preservation actions can be performed successfully onsite by digital archivists and librarians, the lack of a sensitive data DDP service was identified as a significant gap for these institutions. As a result, data are at a high risk of loss because they are usually only stored locally and rarely replicated elsewhere; thus, these data are excluded from services which provide the essential and standards-based components of digital preservation such as geographical distribution. Sensitive data can be found in almost all archives and is prevalent in many cultural heritage organizations but because of the legal and technical complexities involved in preserving such data over a network of providers no existing non-profit DDP network currently provides a HIPAA/FERPA compliant preservation service.[9]

## Methodology

Beyond the main project partners and IMLS, the project team is grateful for the support and participation of more than a dozen institutions. These partners have, among other things, helped us collect use cases for private and sensitive data preservation and helped surface related technical, legal and service model needs and challenges.

The project leads met online roughly every two weeks from September 2019 - January 2020 and then monthly until the end of the grant term. GRA Hesam Andalib from UT iSchool joined the project from September 2019 and met independently with PI Courtney Mumma weekly until he joined the full team meetings from May until the end of his employment term. From September through January 2020, his main duty was to gather data and collect use cases. He also contributed several diagrams of the contributing technical and legal infrastructure.

We convened grant personnel and advisors for a one-day, in-person meeting in Austin, Texas, on December 5th of 2019.[10] The attendees, listed below, included representatives from project partners at archives, libraries, service providers, and supercomputing centers across the United States as well as grant support staff from TDL and a representative from the technical security analytics consulting firm.

---

[9] The Academic Preservation Trust (APTrust) does not have HIPAA/FERPA certification. It uses Amazon commercial services exclusively and does not accept sensitive data unless it is encrypted to standards that meet its depositors' own individual institutional requirements for handling such data. APTrust will allow ingestion if the depositors encrypt it themselves, and APTrust does not hold the keys to decrypt the content.

[10] In-Person meeting (see https://texasdigitallibrary.atlassian.net/wiki/x/AQA4Ow with links to group notes)

- Attendees:
    - Ashley Adair, Digital Archivist, University of Texas at Austin
    - Hesam Andalib, Graduate Research Assistant, Texas Digital Library
    - David Bliss, Digital Processing Archivist, University of Texas at Austin
    - Bill Branan, Senior Engineering Lead, LYRASIS
    - Jaime Combariza, Director, Maryland Advanced Research Computing Center (MARCC), Johns Hopkins University
    - Lea DeForest, Communications Manager, Texas Digital Library
    - Shi Dong, Research Assistant, Northeastern University
    - Chianta Dorsey, University Archivist, University of Texas Southwestern Medical Center
    - Chip German, Program Director, APTrust, University of Virginia
    - Kelly Gonzalez, Assistant Vice President for Library Services, University of Texas Southwestern Medical Center
    - Lauren Goodley, Archivist, Texas State University
    - Ramona Holmes, Associate Director, University of North Texas Health Science Center
    - Chris Jordan, Data Management Group Lead, Texas Advanced Computing Center (TACC), UT Austin
    - Susan Kung, AILLA Archives Manager, University of Texas at Austin
    - Tim Marconi, Director of IT, University of California - San Diego (virtual)
    - Nathaniel Mendoza, Manager, Networking, Security & Operations, Texas Advanced Computing Center (TACC), UT Austin
    - David Minor, Manager, Bio-Med Library, University of California - San Diego
    - Courtney Mumma, Deputy Director, Texas Digital Library
    - Kristi Park, Executive Director, Texas Digital Library
    - Sibyl Schaefer, Chronopolis Program Manager, University of California - San Diego
    - Jen Stone, Principal Security Analyst, SecurityMetrics
    - Alex Suarez, Administrative Associate, Texas Digital Library
    - Danielle Whitehair, Sherlock Project Manager, San Diego Supercomputer Center (virtual)

The project leaders presented updates to the attendees about the project goals and objectives and offered high level descriptions of the current service models available at Chronopolis and TDL as well as those of APTrust and LYRASIS/DuraCloud. Once the stage was set, the meeting focused on collecting input from attendees notably \their own perspectives and use cases concerning private and sensitive data at their institutions and in their varied roles. Additionally, we discussed what was being potentially overlooked in this investigation as well as what roadblocks related to contracting for services they may have experienced at their institutions. In particular, we talked about gaps in the existing DDP technical infrastructure and anticipated problems we will face adapting it to private and sensitive data needs. Most attendees had some level of experience with various service models, so we gathered feedback about elements of good governance including stakeholders, roles, and cost models.

Institutions expressed a need for a DDP service for the private and sensitive data since they are currently holding or planning to accept data with elevated confidentiality requirements.

The partners and the project team grappled with defining what qualified as sensitive and private content. There was little clarity about the level of protection required across different individual units within an institution, between institutions, from state to state and across national boundaries. There are complexities in determining the control and ownership of content and most of the partner institutions indicated a lack the capacity and/or resources to properly determine the extent of private and sensitive data at risk in their possession. Some do not accept transfer of any content that may contain such data since they do not have the means to protect it or to provide properly mediated access to it.

After the in-person meeting, the core grant team continued to gather data about legal issues, collect use cases and existing contracts and other legal binds, document areas of concern, research costs, and illustrate current workflows. Since January 2020, we have been analyzing all of the data gathered and information from the in-person meeting.

Unfortunately, due to the COVID-19 pandemic it became impossible to travel to conferences where we had intended to engage in more discussion with colleagues about the directions of the project. The pandemic also decreased staffing availability as reacting to the current state of affairs became the top priority.

Fortunately, the team had always intended for the final wrap meeting to be virtual, and on Friday, August 21, 2020, we convened grant personnel and advisors for a three hour virtual wrap meeting. Attendees, listed below, included representatives from project partners at archives, libraries, service providers, a university privacy officer, supercomputing centers across the United States, as well as grant support staff from TDL. Not all of the same attendees from the in-person meeting were able to attend, but we were able to add new partners from government archives, museums and different representatives from institutions which had been represented before and/or interviewed by the project team.

- Attendees:
    - Ashley Adair, Digital Archivist, University of Texas at Austin
    - Hesam Andalib, Graduate Research Assistant, Texas Digital Library
    - David Bliss, Digital Processing Archivist, University of Texas at Austin
    - Bill Branan, Senior Engineering Lead, LYRASIS
    - Sandeep Chandra, Executive Director for Sherlock, San Diego Supercomputing Center
    - Jaime Combariza, Director, Maryland Advanced Research Computing Center (MARCC), Johns Hopkins University
    - Bradley Daigle, Digital Initiatives Librarian and AP Trust Content and Strategic Expert, Academic Preservation Trust
    - Lea DeForest, Communications Manager, Texas Digital Library
    - Chianta Dorsey, University Archivist, University of Texas Southwestern Medical Center
    - Chip German, Program Director, APTrust, University of Virginia
    - Kelly Gonzalez, Assistant Vice President for Library Services, University of Texas Southwestern Medical Center
    - Lauren Goodley, Archivist, Texas State University

- Heather Greer Klein, DSpace Product Manager, LYRASIS
- Ramona Holmes, Associate Director, University of North Texas Health Science Center
- Chris Jordan, Data Management Group Lead, Texas Advanced Computing Center (TACC), UT Austin
- Susan Kung, AILLA Archives Manager, University of Texas at Austin
- Meg McAleer, Senior Archives Specialist, Library of Congress
- Isabel Meyer, DAMS Branch Manager, Office of the Chief Information Officer, Smithsonian
- David Minor, Manager, Bio-Med Library, University of California - San Diego
- Courtney Mumma, Deputy Director, Texas Digital Library
- Francis Park, Historian, Joint History and Research Office, Joint Chiefs of Staff
- Kristi Park, Executive Director, Texas Digital Library
- Pegah Parsi, Campus Privacy Officer, University of California, San Diego
- Sibyl Schaefer, Chronopolis Program Manager, University of California - San Diego
- Lydia Tang, Special Collections Archivist, Michigan State University

During the meeting, the project leads reviewed the objectives and work completed to date. We polled the attendees to find out whether they were engaged at all with any kind of DDP network, finding that 15% of attendees were depositors, 35% represented a service provider and 60% were not at all engaged in any DDP networks. Next, the project leads reviewed three service options they had devised based on the information gathered so far, including summary information about technical and service requirements including costs. After presenting the three potential solutions, the team polled the attendees about which of the services their institution might participate in. This report will discuss the findings of that poll later in the section describing the various service options.

Before breaking into group discussions of the service options, the project leads reviewed the criteria and legal binds between parties and other service model considerations for each of the service model options. In the breakout groups, 3-4 attendees discussed barriers and advantages of any one of the service options with a project representative. Then, they considered together what capacity building activities might be needed at their institutions to be ready for a DDP for sensitive data. The meeting concluded by sharing summaries of the breakout groups' comments and a discussion about the next steps, including further dissemination and implementation plans.

Within the confines of COVID-19 restrictions, the team has been able to present its methods and findings via SAA, an OCLC Research webinar, NDSA's DigiPres 2020, CNI Spring 2020, WeMissiPres Virtual unConference, and TDL member forums and groups. With the publication of this report, we hope to engage with more interested parties and continue to interrogate community needs.

# Use cases for sensitive data in libraries and archives

In order to assess current practice and establish whether there were gaps in service that needed to be filled for members and aligned institutions, and to test our assumptions about the need for this service and see if they are valid, UT iSchool Graduate Research Assistant Hesam Andalib interviewed representatives from four TDL member institutions as well as two UCSD units. The project team also collected use cases from the in person partner meeting which took place in Austin in late 2019, as well as from a few other organizations we engaged with over the course of the project. Ultimately, nine institutions shared their current private and sensitive data strategies with our team.

In each of these discussions, we asked if institutions were aware of sensitive data justifying a high level of preservation either held by their organization or within their organizational collecting purview. In some cases, institutions suspect that there is such data, but haven't done the level of assessment or appraisal necessary to quantify the problem sufficiently. In other cases, the assessment of sensitive content has been completed and a decision has explicitly been made not to bring it into the custody of the library or archives. Use cases demonstrating this behavior give one or more of the following reasons as examples: limited resources to manage the content, unclear authority to manage it properly, and a dearth of places to keep it.

## Types of private and sensitive data referenced in the use cases

The following is a list of the types of sensitive and private data discussed as part of the use cases described below. Note that this project has deliberately excluded all classified secret, top secret and confidential data under the United States Government Classification System.[11]

- PHI and other health data governed by HIPAA as well as clinical data used in teaching and historical medical records (3 institutions)
- Any records containing PII including email and other correspondence, research data, digitized materials, documents, manuscripts, maps, images and audiovisual recordings (9 institutions)
- Student records governed by FERPA (3 institutions)
- Human Rights archives and accounts of personal trauma (1 institution)
- Unprocessed and under-described collections (9 institutions)
- Commerce-related restricted data (2 institutions)

Other use cases mentioned in less detail included:

- Potentially patentable data. Universities sometimes designate data according to its patentability or potential for commercialization, etc.
- Sensitive data that may appear in unexpected places in unprocessed material, such as architectural collections.

[11] Issued by President Barack Obama in 2009, Executive Order 13526 replaced earlier executive orders on the topic and modified the regulations codified to 32 C.F.R. 2001. "Executive Order 13526 - Classified National Security Information". Information Security Oversight Office of The National Archives. Retrieved January 5, 2010.

- National hazards engineering data. Data that logs coordinates of house damage or photographic evidence of housing disasters, for example, are considered sensitive as an individual's personal items and living spaces are displayed.

While both TDL and UCSD have a functioning service model for their own current DDP network, the team wanted to gather input from partners about their own preferences and experiences as participants and leaders of similar networks. Both TDL and UCSD, as well as several other project partners, were involved in the Digital Preservation Network, a DDP service that failed for numerous reasons, many of them related to governance and the cost model.[12] With this in mind, we asked our stakeholders to help us identify the elements of good governance for this type of service. Their answers are as follows:

- Clear vision, mission, roles and responsibilities
- Centralized decision-making with stakeholder consultation
- Diverse representation of institution types and practitioners
- Responsiveness to legal fluctuation and jurisdictional differences
- Transparent financial reporting
- Succession planning
- Clear role for data owner(s) (ownership, copyright, access controls)
- Node staff representation
- Standards-driven
- Collections-driven (different content needs across various collection-types in libraries and archives)
- Encouragement of collaboration among stakeholders/members
- Open communication

## Data Stakeholders and Data Ownership

One of the most engaging discussions at the in-person meeting was centered around identifying potential data stakeholders and questions concerning which of those stakeholders were legally and/or morally ‘owners’ of the data.

The various stakeholders identified include:

- Depositors: the people or organizational units placing the data in the custody of the archives or libraries.
- Preservation staff: repository managers, archivists, librarians, developers and service providers.
- Human subjects in the data: For example, there are the people in photographs or represented in the health records, the relatives of persons in the data (especially in health records), entire communities represented by data, and the original creator of data (ie doctor, researcher).

[12] Pcolar, D. (2019, October 10). Digital Preservation Network (DPN) Final Report. https://doi.org/10.17605/OSF.IO/MD9YK

- External stakeholders: administrators, academic Deans and Directors, EVPs for health system and research for clinical and research data, university presidents for overall support/approval, CIOs and other executive leadership.
- The institutions themselves

For the attendees, ownership seemed to be the factor most frequently used to drive privacy and sensitivity decisions. Without knowing who owns the data, it can be hard to make good decisions about digital preservation. Determination of ownership is central to any discussion of sensitive data preservation, because that determination has consequences regarding the right to set aside, preserve and provide access over time as well as the ultimate act of ownership: destruction. Determination of ownership varies in different contexts. For instance, according to our consultants from SecurityMetrics, in the US, ownership of data is largely driven by commerce, whereas in the EU, it is primarily driven by a privacy imperative. Archivists and librarians are data custodians guided by local classification and regulations which decide how and when to apply decisions made based on ownership. US capitalism has a significant impact on this project and on private and sensitive data kept in libraries, especially with regard to the risk of liability a library takes when ownership and rights are unclear.

When considering privacy and sensitivity, the group also highlighted the need to consider that future users of preserved sensitive content might be operating under different regulations that put the data at greater risk in the future. Because of this particular risk, those managing acquisitions would be wise to include parameters about deletion, access, and de-identification in the legal deeds of gift, if possible.

## Analysis and deeper understanding

In addition to analyzing the information gathered from users and reviewing the research and commentary from our contractors, Sibyl Schaefer enrolled in and completed the requisite training to become a Certified HIPAA Privacy Security Expert (CHPSE). This training, offered by the Supremus Group, LLC is an intensive 22-hour HIPAA course outlining topics such as the application of the HIPAA security rule as it relates to the security of PHI, identifying technical or electronic threats to the healthcare enterprise and the technology available to reduce or prevent those threats, advanced training in the topics of administrative, physical, and technical safeguards and how to develop policies and procedures to describe those safeguards and address larger risk management strategies.

## Related projects

During the course of our investigations and analysis, we discovered two ongoing related projects. First, during the 2020 DLF Forum, which took place online November 9-10, 2020, there was a presentation[13] by John Bowers, Jack Cushman, Jayshree Sarathy, and Jonathan Zittrain

[13] John Bowers, Jack Cushman, Jayshree Sarathy, and Jonathan Zittrain. "'Time Capsule' Archiving Through Strong Dark Archives (SDA)", 2020 Virtual DLF Forum: https://youtu.be/tVhcTfxj7IM. Last accessed January 20, 2021.

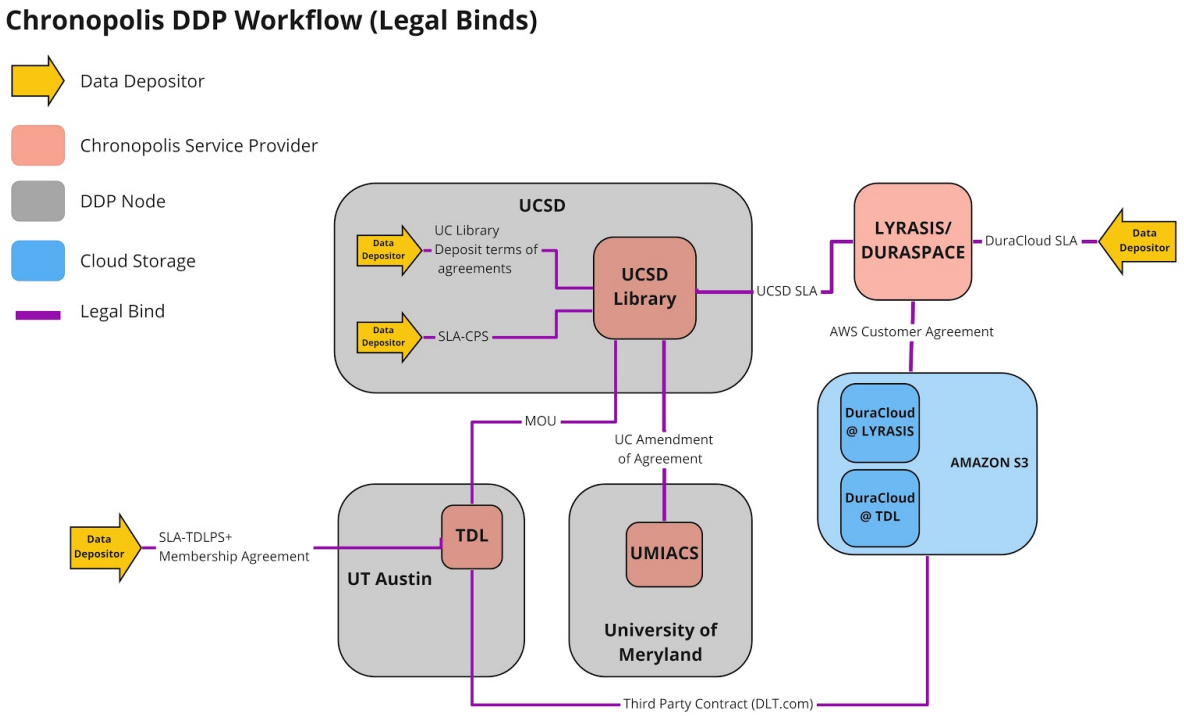
that recognized that sensitive digital artifacts pose new challenges for delayed-release archiving. To protect our historical record for future generations, they proposed Strong Dark Archives (SDA), a blended administrative and technical protocol for securing delayed-released archival materials among networks of libraries. Their model used distributed secret keys across notes to encrypt content and enable coordinated release, which would require mediated cooperation across the nodes to access or make changes.

The second project, by Alex Garnett and Jin Zhang, was also presented at the DLF Forum and is an initiative to add optional zero knowledge encryption to the Federated Research Data Repository in Canada. This means the Repository administrators will never be able to access the data directly, nor would any malicious intruder to the system. Instead, all data deposited into the Repository will be encrypted with keys deposited into a separately managed platform, using Hashicorp's Vault software.[14]

## Legal binds that enable the Chronopolis service

There are many connections in the active Chronopolis and TDL partnerships which require legal agreements. The connections you see illustrated below include software licenses (for example, those which are in place for DuraCloud, Chronopolis, and Amazon Web Services; they also include contracts, service Level Agreements (SLAs) and/or Memoranda of Understanding (MOUs) between the service provider and storage node (like TDL with TACC and Chronopolis with NCAR); between service provider and depositor (like UCSD and TDL with their members and community depositors); and between 2 service providers (like those agreements between Chronopolis and TDL as well as between LYRASIS/DuraCloud and Chronopolis).

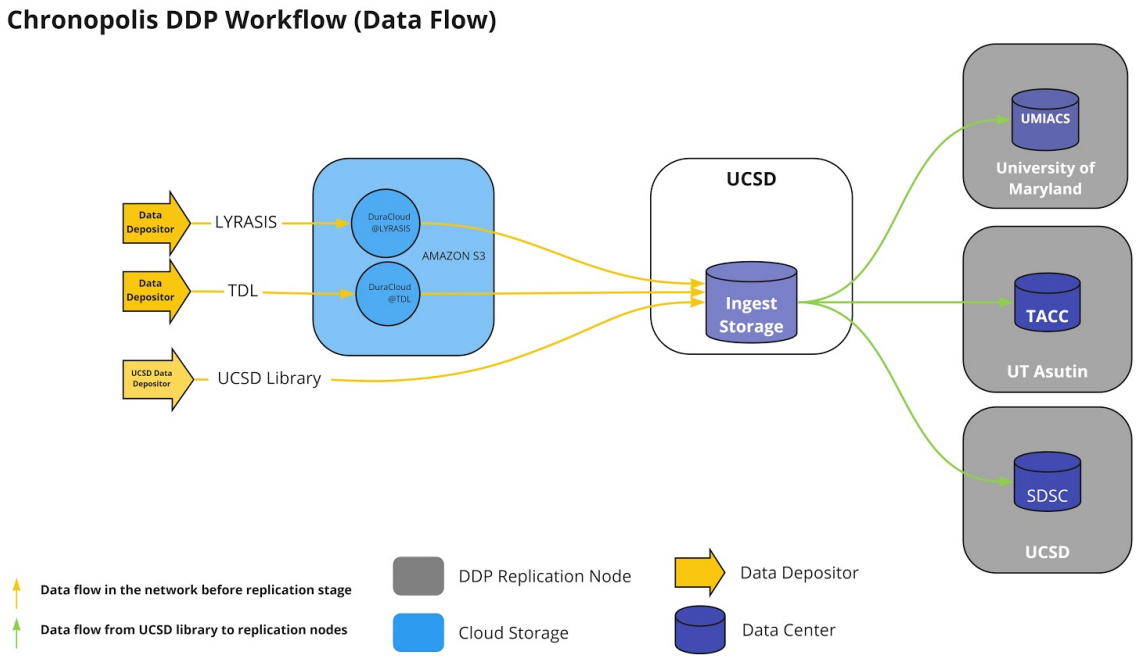
[14] Portage Network, SFU Working Towards Zero Knowledge Encryption of Sensitive Data in FRDR. https://portagenetwork.ca/news/sfu-working-towards-zero-knowledge-encryption-of-sensitive-data-in-frdr/ (accessed February 5, 2021)




The project team recognizes that the complexity of this system will be compounded by regional and jurisdictional boundaries, local policies and regulations, and audit needs. While the project team originally intended to provide templates for enabling legal agreements for a DDP system for private and sensitive data, it became clear through our investigations that these complexities prohibit the usefulness of such templates.

### Current Chronopolis Technical Infrastructure

Should TDL and UCSD move ahead with their own DDP for private and sensitive data, its design ideally would be consistent with the current architecture to reduce costs and complexity. The current Chronopolis DDP data flow is illustrated below. Data depositors at LYRASIS and TDL send their data to Chronopolis ingest storage via DuraCloud instances mounted in Amazon S3, while UCSD depositors ingest directly into UCSD's Chronopolis ingest storage. From this storage, data is replicated to the UMIACS, TACC and UCSD replication nodes for preservation.




## Understanding Private and Sensitive data requirements

### SecurityMetrics Analysis

To highlight the questions that need to be answered for designing a DDP network for private and sensitive data, we consulted with the Principal Security Analyst at Security Metrics who reviewed our legal documents. In their report to us,[15] the following questions were raised:

- It is critical to understand that a blanket classification of all sensitive data may increase both the legal and technical burdens applied to each set of data deposited.
- Data may fall under different privacy laws. Privacy laws come with organizational (non-technical) requirements made up of policies, procedures, and training.
  If all sensitive data needs to be protected in the same way, consideration should be given to a privacy law crosswalk approach that consolidates the language in some way.
- Decisions include authorizing access, determining the rules for modification or deletion, procuring third-party audits, etc. It is important to understand who is able to make decisions about sensitive information in each case.
- Sensitive data may fall under laws that require it to be maintained for a certain length of time. One reason to establish roles and responsibilities is to know who will absorb costs if the original payer is unable to maintain the information.

### Data Classification Standards and Implementation Requirements

Many of the questions that came up in the SecurityMetricsreport can be answered by local policies and regulations. For instance, data held in any of TDL's storage systems located at UT

[15] Jen Stone (2020), Preserving Sensitive Data in Distributed Digital Storage Networks, Security Metrics

Austin, including TACC are governed by the UT Information Security Office's Data Classification Standard.[16] The University of California also provides assistance in classifying information based on confidentiality and integrity requirements in the "UC Protection Level Classification" Guides.[17]

Presumably, institutional depositors have their own such local policies and regulations regarding data classification, including details about storage succession and decision-making. Those policies all likely have procedural recommendations and requirements. As a service provider, adherence to policies throughout the system will have to be strictly codified and enforced.

### HIPAA requirements[18]

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) required the Secretary of the U.S. Department of Health and Human Services (HHS) to develop regulations protecting the privacy and security of certain health information. To fulfill this requirement, HHS published what are commonly known as the HIPAA Privacy Rule and the HIPAA Security Rule. The Privacy Rule, or *Standards for Privacy of Individually Identifiable Health Information*, establishes national standards for the protection of certain health information. The *Security Standards for the Protection of Electronic Protected Health Information* (the Security Rule) establish a national set of security standards for protecting certain health information that is held or transferred in electronic form. The Security Rule requires covered entities to maintain reasonable and appropriate administrative, technical, and physical safeguards for protecting PHI.
The HIPAA Privacy regulations require health care providers and organizations, as well as their business associates, to develop and follow procedures that ensure the confidentiality and security of PHI when it is transferred, received, handled, or shared. This applies to all forms of PHI, including paper, oral, and electronic, etc.[19]

### Covered Entity

Individuals, organizations, and agencies that meet the definition of a "covered entity" under HIPAA must comply with the Rules' requirements to protect the privacy and security of health information and must provide individuals with certain rights with respect to their health information. Covered entities are defined in the HIPAA rules as (1) health plans, (2) health care clearinghouses, and (3) health care providers who electronically transmit any health information in connection with transactions for which HHS has adopted standards. Researchers are covered entities if they are also health care providers who electronically transmit health information in

[16] This standard serves as a supplement to the Information Resources Use and Security Policy, which was drafted in response to Texas Administrative Code 202 and UT System UTS-165. Adherence to the standard will facilitate applying the appropriate security controls to university data.
https://security.utexas.edu/policies/data_classification
[17] This guide is part of the UC's revised and updated Electronic Information Security Policy (IS-3) that aims to protect user confidentiality; to maintain the integrity of all data created, received or collected by UC.
https://security.ucop.edu/policies/institutional-information-and-it-resource-classification.html
https://security.ucop.edu/files/documents/uc-protection-level-classification-guide.pdf
[18] Most of the information in this section is retrieved from US Department of Health and Human Services website-Health Information Privacy https://www.hhs.gov/hipaa/index.html
[19] From Department of Health Care Services
https://www.dhcs.ca.gov/formsandpubs/laws/hipaa/Pages/1.00WhatisHIPAA.aspx

connection with any transaction for which HHS has adopted a standard. Because UT Austin and UCSD both provide health care services to the general public, they are considered covered entities and thus must comply with HIPAA. This is the case for many of the universities we partnered with on this project.

### Business Associate

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. By law, the HIPAA Privacy Rule applies only to covered entities. However, most health care providers and health plans do not carry out all of their health care activities and functions by themselves. Instead, they often use the services of a variety of other persons or businesses. The Privacy Rule allows covered providers and health plans to disclose protected health information to these “business associates” if the providers or plans obtain satisfactory assurances that the business associate will use the information only for the purposes for which it was engaged by the covered entity, will safeguard the information from misuse, and will help the covered entity comply with some of the covered entity’s duties under the Privacy Rule.
If a covered entity engages a business associate to help it carry out its health care activities and functions, the covered entity must have a written Business Associate Agreement, or BAA or other arrangement with the business associate that establishes specifically what the business associate has been engaged to do and requires the business associate to comply with the Rules’ requirements to protect the privacy and security of protected health information. If TDL and UCSD did form a HIPAA- compliant DDP network, we would be considered Business Associates and need the appropriate BAAs in place.

### Encryption

The project team recognizes that encryption is not best practice for digital preservation, but that in some cases institutions consider it their only choice among few alternatives. While encryption protects PHI by significantly reducing the risk of the information being viewed by unauthorized persons, such protections alone cannot adequately safeguard the confidentiality, integrity, and availability of PHI as required by the Security Rule. Encryption does not maintain the integrity and availability of the PHI, such as ensuring that the information is not corrupted by malware, or ensuring through contingency planning that the data remains available to authorized persons even during emergency or disaster situations. Further, encryption does not address other safeguards that are also important to maintaining confidentiality, such as administrative safeguards to analyze risks to the ePHI or physical safeguards for systems and servers that may house the PHI.

Also it is important to know that storing encrypted PHI and lacking the key to encrypted data does not exempt a Cloud Service Providerfrom business associate status and associated obligations under the HIPAA Rules. An entity that maintains PHI on behalf of a covered entity (or another business associate) is a business associate, even if the entity cannot actually access the PHI. There is an expectation that data will be encrypted and that there will be security breach notifications if it is not.

### When is Private Health Information not protected by HIPAA?

The HIPAA Privacy Rule protects the individually identifiable health information about a person for 50 years following their death. During the 50-year period of protection, the personal representative of the decedent (i.e., the person under applicable law with authority to act on behalf of the decedent or the decedent's estate) has the ability to exercise the rights under the Privacy Rule with regard to the decedent's health information, such as authorizing certain uses and disclosures of, and gaining access to, the information. The Privacy Rule permits a covered entity to disclose the relevant protected health information of the decedent to family members or other persons involved in the individual's health care or payment for care prior to the individual's death, but who are not personal representatives.. Tus, the 50-year period of protection balances the interests of surviving relatives with the need for archivists, biographers, historians, and others to access records on deceased individuals for historical purposes.

### FERPA requirements

The Family Educational Rights and Privacy Act of 1974[20] is a United States federal law that governs the access to educational information and records by public entities such as potential employers, publicly funded educational institutions, and foreign governments. The Act serves two primary purposes:

1. It gives parents or eligible students more control of their educational records
2. It prohibits educational institutions from disclosing "Personally Identifiable Information in education records" without written consent.

Any public or private school or any state or local education agency must comply with FERPA rules. FERPA prohibits the disclosure of a student's "protected information" to a third party. For purposes of FERPA, a "third party" includes any individual or organization other than the student or the student's parent(s). With respect to third parties, even if the initial disclosure of protected information is permissible, FERPA limits the subsequent disclosure of the information by the third party. As such, once an educational institution discloses protected information to a third party, it must ensure that the third party does not itself improperly disclose the information in violation of FERPA.

FERPA classifies protected information into three categories: educational information, personally identifiable information, and directory information. The limitations imposed by FERPA vary with respect to each category. Personally identifiable information can only be disclosed if the educational institution obtains the signature of the parent or student (if over 18 years of age) on a document specifically identifying the information to be disclosed, the reason for the disclosure, and the parties to whom the disclosure will be made. Failure to comply with these requirements will result in a violation of FERPA.
Directory information is defined as "information contained in an education record of a student that would not generally be considered harmful or an invasion of privacy if disclosed.[21]

[20] FERPA. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html (accessed January 21, 2021)
[21] FERPA Primer: The Basics and Beyond. https://www.naceweb.org/public-policy-and-legal/legal-issues/ferpa-primer-the-basics-and-beyond/ (accessed January 21, 2021)

### When are Educational Records not protected by FERPA?

FERPA rights of eligible students lapses or expires upon the death of the student. Therefore, FERPA would not protect the education records of a deceased eligible student (a student 18 or older or in college at any age) and an educational institution may disclose such records at its discretion or consistent with State law. However, at the elementary and secondary levels, FERPA rights do not lapse or expire upon the death of a student because FERPA provides specifically that the rights it affords rest with the parents of students until that student reaches 18 years of age or attends an institution of postsecondary education. Once the parents are deceased, the records are no longer protected by FERPA.[22]

## Private and Sensitive Data Service Models

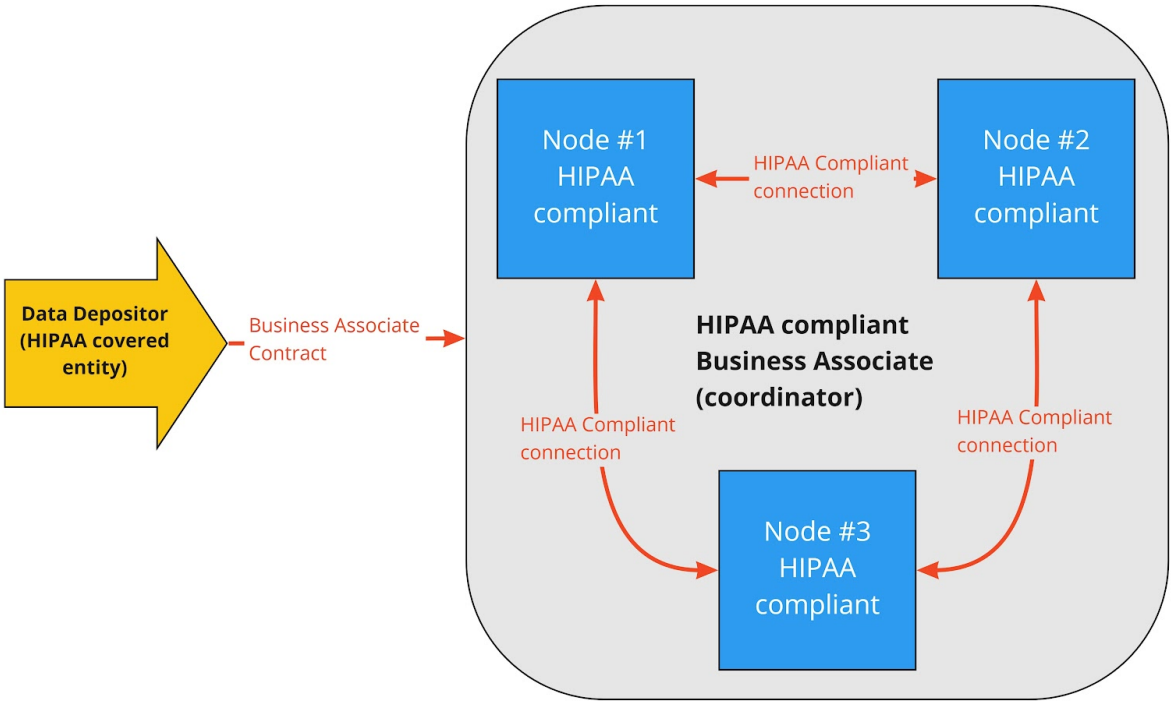
### Fully HIPAA compliant

#### Minimal demonstration of a HIPAA Compliant Service

The diagram below illustrates the very minimum, basic configuration and requirements of a HIPAA-compliant DDP network, and it includes the essential roles of ‘business associate’ and ‘covered entity’. Understanding these two roles, as defined in a prior section, 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. The basic conceptual model below shows any data depositor who qualifies as a covered entity would submit their content into a distributed digital preservation environment managed by a business associate via a transaction that is governed by a BAA.. The business associate manages every connection that moves data through the network, and monitors the storage distributed across at least three nodes. Those nodes are managed in a facility in alignment with HIPAA requirements.

[22] Does FERPA Protect Education Records of Students that are Deceased? https://studentprivacy.ed.gov/faq/does-ferpa-protect-education-records-students-are-deceased (accessed January 21, 2021)

**Basic Conceptual model for HIPAA/FERPA compliant DDP network**




The model is deceptively simple. The project team took great pains to understand the connections, data movement and legal binds required to achieve their current digital preservation network models in Chronopolis. As illustrated and described previously in this report, the details of the Chronopolis model complicates the application of the above representation, as there are several subcontracts that would need to be in place where HIPAA rules governing the activities and responsibilities of a business associate (UCSD, in this case) would have to be applied. It is for this reason, in fact, that the project team has formulated these three potential plans to accommodate a range of private and sensitive data, not all of which are fully HIPAA compliant.

Most institutions who contributed their use cases and feedback to this planning grant 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 fortifying readiness in the beginning stages of the process, the need for DDP storage seems far off for all except a very small group of institutions.

For that small subset of institutions who are ready to move forward with PHI and PPI storage, there are extensive technical and legal connections required to enable our current DDP services

and which will be needed to move forward with any new DDP for sensitive data options. The complexity of these requirements is compounded by audit requirements for partners as well as new nodes across different jurisdictions and boundaries. We also face the problem of an ever evolving legal environment. There are implications for DDPs when data is received under one expectation of privacy that changes as the legal environment does over time.

Costs differ depending on the storage facility, but in all cases, they are still more than “regular” digital preservation 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. This is incompatible with the way that most archives manage their aggregations. Archives are described in hierarchical groupings and not at the individual item level. Item-level processing of massive collections to find private and sensitive data is out-of-reach for most archives and special collections, resulting in storage of collections likely to contain such data in the aggregate for placement in DDP sensitive data storage.

For those reasons, the grant partners are drafting recommendations for multiple service options, described in detail below. First, UCSD and/or TDL could independently offer a fully HIPAA-compliant single storage node solution to their respective partners by leveraging their existing partnerships with SDSC and TACC. These two nodes could also be connected to start forming a full DDP network (which would have a minimum of three nodes). A third option is to create a HIPAA-like DDP network, meeting all the requirements but without the costly audit.

### Single service node

A first option is to create a single service node under the condition that an institution has at least one but preferably two of its own geographically distributed and HIPAA-compliant digital preservation storage options.

The Texas Digital Library Digital Preservation Service 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 for two reasons. First, Chronopolis is a dark archive and is designed to provide copies of content in emergency or disaster scenarios, not for regular business operations. Second, institutions using the Amazon options TDL provides should have a second or third node available outside of Amazon to avoid egress costs for immediate access. 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 rely on their IT departments to manage two locally distributed digital preservation copies by provisioning backup areas for their materials over which they have control. Relationships between archives or libraries and their associated information technology can be complicated, but TDL offers consulting and strategy support to Digital Preservation Services members to assist in technological support and storage provisioning as well as 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 BAAs are in place appropriately. UCSD would not offer a single node service beyond their own institution in order to avoid taking on added risks.

### Fully Distributed via DDP service

Currently, UCSD and TDL partner with data centers at SDSC and TACC, each of whom provide HIPAA-compliant storage. If a third node with similar access to a data center with HIPAA storage joined the two nodes, 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 if the agreements were updated to BAAs, 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.All parties would likely need to consult with their respective privacy officers and legal departments to ensure that the institutions felt confident in their administrative and technical capacity to take on the risks associated with managing PHI and PII.

### Fully Distributed HIPAA-like DDP

Because of the costs associated with HIPAA audits 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 detailed in the discussion of the use cases .

Such an offering could still leverage HIPAA storage at SDSC (Sherlock)  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. DDP services like Chronopolis already provide a high level of security during ingest and replication. However, the HIPAA-like service option would need to include clear boundaries that exclude PHI falling under the HIPAA rules due to a risk of liability should that data get stored in the system. Over time, Chronopolis and TDL could work towards a two node HIPAA-eligible network as we look to secure a third node. Eventually, we could become fully HIPAA compliant across the entire infrastructure.

## Encryption Requirements

If data is encrypted, the keys can be stored either by the depositor or by the provider. Either way, as discussed in the section describing HIPAA requirements, the service provider is obligated as a business associate to adhere to standards for PHI. For the purposes of this report, that means that providers offering either of the fully-HIPAA compliant options could make their own decisions about whether they would also support key handling or whether they would leave it to the depositor. None of the service options, including the fully HIPAA-compliant ones, require encryption in motion or at rest. It’s very complicated, but basically, encryption is required for HIPAA if it’s reasonable and appropriate to encrypt.

> "The covered entity must decide whether a given addressable implementation specification is a reasonable and appropriate security measure to apply within its particular security framework. For example, a covered entity must implement an addressable implementation specification if it is reasonable and appropriate to do so, and must implement an equivalent alternative if the addressable implementation specification is unreasonable and inappropriate, and there is a reasonable and appropriate alternative."[23]

The use of encryption has been controversial in digital preservation good practice, but it is generally agreed that encryption blocks essential maintenance and monitoring activities. The HHS documentation of the Security Rule goes on to explain the requirements should an entity decide not to encrypt.

> "This decision will depend on a variety of factors, such as, among others, the entity's risk analysis, risk mitigation strategy, what security measures are already in place, and the cost of implementation. The decisions that a covered entity makes regarding addressable specifications must be documented in writing. The written documentation should include the factors considered as well as the results of the risk assessment on which the decision was based."

Therefore, if a DDP service provider chooses to provide HIPAA-compliant storage, there would need to be documentation of the choice not to encrypt as well as the rigorous alternative methods that the DDP system has implemented in its architecture to secure the data in alignment with digital preservation best practice. It's likely that medical libraries and health science centers would reject an unencrypted option. Alternately, services should dedicate resources to investigation and planning for long-term key and encryption management to accommodate satisfaction of HIPAA guidance, at least unless and until a new way of securing private and sensitive data comes along to replace encryption that better aligns with digital preservation goals.

## Summary of Lessons Learned

The project team has uncovered key lessons which will inform any service providers interested in developing DDP services for private and sensitive data. 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. While discussing the use cases, the partners and the project team struggled with defining what qualified as sensitive and private content. There is also little understanding about the levels of protection required across different individual units, between institutions, from state to state and across national boundaries. There are complexities in the determination of control and ownership of content, 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. Further, some of the partners refuse to acquire or accept transfer of any content that may contain such data since they know they do not have the

[23] HIPAA Security Rule. https://www.hhs.gov/hipaa/for-professionals/security/index.html

means to protect it or to provide properly mediated access to it. 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.

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. They are, however, added complexity which only becomes more so when you add audit requirements for partners and new nodes across different jurisdictions and boundaries. With complexity comes added costs. Those costs differ depending on the storage facility, as shown in the pricing differences between TACC and UCSD.

It is common practice to deposit unprocessed material to secure it as a sort of “triage,” with the intention to process materials at a later date. This practice would also result in higher costs at least until the materials can be reviewed and properly appraised, and even when those set aside for retention are identified, there will be additional costs to remove the deselected data from any DDP system.

## Next Steps

When thinking about how to move ahead with private and sensitive data DDP service provision, we need to assess whether there *is* actually HIPAA-covered data at a state in the digital preservation lifecycle workflow that makes it ready to move into digital preservation storage. If so, is it enough to justify a service? If it is enough to justify a service, which service model is the most reasonable or appealing? We polled our partners in the wrap meeting and the HIPAA-like and the Single-node offerings were most appealing, simply because they could be offered at a reduced cost. TDL and UCSD will need to make decisions about which service offering is the most feasible. We could start with a one node offering and expand as the market expands. Over time, we could work towards a two node HIPAA-eligible network as we look to secure a third node. Eventually, we could become fully HIPAA compliant across the entire infrastructure.

For a DDP service to be successful, it needs to have a healthy number of depositors buying into the system to support it. What we found in this project is that the market is not yet developed enough to support this type of service offering effectively. Many of the institutions we worked with throughout the project are not yet engaged with DDP networks for their typical digital preservation needs, much less for more complicated needs associated with private and sensitive data.

Building capacity to manage private and sensitive data so institutions are ready for storage solutions is a foundational gap we discovered over the course of our investigations. Do potential depositors have digital preservation expertise and if not, can they manage the resources needed to get good training? Once they are trained up and have policies and procedures in place, do they have the funding to undertake the extra layers of processing necessary for digital preservation and/or for sensitive data? How will they acquire access to the skills necessary to mitigate risk involved with handling HIPAA and FERPA content?

As a community of digital preservation practitioners, we concluded that it is necessary to take a step back to fully consider these readiness questions  and recognize the need for proper preparation of collecting institutions to acquire private and sensitive data. Readiness remains a

significant obstacle across all types of digital preservation, not just for private and sensitive data. And a “build it and they will come” approach is impractical given the legal, financial, organizational, and other complexities of building a fully distributed digital preservation network that meets HIPAA regulatory requirements. In the meantime, the project partners will continue to evaluate possibilities for single-node preservation or “HIPAA-like” storage that may meet certain needs of stakeholders discovered in our research.