sciendo

Position Article **Open Access**

Beth A. Plale*, Eleanor Dickson, Inna Kouper, Samitha Harshani Liyanage, Yu Ma, Robert H. McDonald, John A. Walsh, Sachith Withana

Safe Open Science for Restricted Data

https://doi.org/10.2478/dim-2019-0005
received September 3, 2018; accepted February 24, 2019.

Abstract: Open science is prompting wide efforts to make data from research available for broader use. However, sharing data is complicated by important protections on the data (e.g., protections of privacy and intellectual property). The spectrum of options existing between data needing to be fully open access and data that simply cannot be shared at all is quite limited. This paper puts forth a generalized remote secure enclave as a socio-technical framework consisting of policies, human processes, and technologies that work hand in hand to enable controlled access and use of restricted data. Based on experience in implementing the enclave for computational, analytical access to a massive collection of in-copyright texts, we discuss the synergies and trade-offs that exist between software components and policy and process components in striking the right balance between safety for the data, ease of use, and efficiency.

Keywords: open science, computational analysis, HathiTrust, capsule framework, restricted data, security, safe open science

***Corresponding author: Beth A. Plale**, Department of Intelligent Systems Engineering, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA, E-mail: plale@indiana.edu
Eleanor Dickson: HathiTrust, University of Michigan, Ann Arbor, MI, USA
Inna Kouper: Department of Informatics, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
Samitha Harshani Liyanage: HathiTrust Research Center, Indiana University, Bloomington, USA
Yu Ma: HathiTrust Research Center, Indiana University, Bloomington, USA
Robert H. McDonald: University Libraries, University of Colorado Boulder, Boulder, CO, USA
John A. Walsh: Department of Information and Library Science, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
Sachith Withana: Department of Intelligent Systems Engineering, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA

1 Introduction

Data have long served as a basis for discovery. The naturalist, Charles Darwin, formed his theories around evolution in the mid-19th century after spending five years aboard the vessel, HMS Beagle, collecting data on plants and animals from around the globe. As computers made their way into science in the 20th century, data itself increasingly were born digital. As computer disks grew in size and data sources (sensors, social media, etc.) grew in number, digitized data grew to a sufficient critical mass to form the basis for new theories of science enabled by sophisticated software for analyzing the data. This scale in growth in the volume and variety of data in science simultaneously made existing forms of conveying data increasingly inadequate whether in tables in peer review publications or in supplemental materials.

Open science is a global effort to make data emerging from scientific research available for broader research and societal and economic uses. Inherent in open science is a recognition of the intrinsic value of research data.[1] Open science is frequently mistaken for open access. Open access data are freely available, free of cost, or other barriers to its access. Open science allows for limited forms of data availability, particularly for data that may need protections of privacy and intellectual property, protection of human research participants, etc.

Much data resulting from externally funded research can be made available, but some data simply cannot or will ever be completely and freely open. Data should be made open to the fullest extent possible (open access, open use, open license), but there are important cases where controls on the data must be in place. For this latter data, an accommodating principle coined by the European Union Horizon 2020 program document on FAIR Data

1 "Research data is the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples), trade secrets, personnel and medical information, etc.." Federal Register 2 CFR 200.315(e)(3)

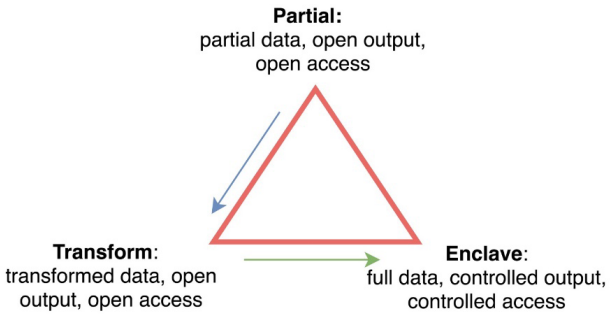

Figure 1. Forms of data availability on the spectrum between fully open access and fully hidden.

Management is that data should be *open as possible* and *closed as necessary* (Rabesandratana, 2013).

The options for carrying out open science on restricted data are limited. Informal sharing arrangements, where a researcher shares a dataset with a colleague under the assurance that the dataset will not be shared further, have existed for quite some time. These arrangements work where a high level of mutual trust exists between the sender of the dataset and its receiver. In the more prevalent case, where trust is low or compliance is an issue (e.g., Institutional Review Boards), the options for making restricted data available are more limited as shown in Figure 1. In each of the three options depicted in the figure, an analyst is assumed to interact with digitized data through software tools. These tools may be as simple as a spreadsheet or a PDF reader or as complex as a deep learning algorithm.

The first form of data availability makes partial information available, labeled as *partial* in the figure. The metadata of a dataset may be made visible (shared) while the dataset itself is not, for instance. A portion of the data/metadata is made open. We assume for purposes of this paper that metadata and data are distinct from one another and that there are no restricted data present in the metadata. We acknowledge that the distinction between data and metadata can be complex and that our characterization may oversimplify in some cases. However, the partial form of sharing encompasses cases where only a portion of meta-information about the restricted data can be openly shared. Sharing of partial data/metadata limits the reuse of the data, but nevertheless facilitates discovery and could even be sufficient on its own. For instance, information about the very existence of a dataset can be used to ascertain the concentration of research in an area.

The second form of data availability is labeled *transform*, where a dataset or portions of a dataset undergo some form of transformation be it derived, aggregated, anonymized, or synthesized to obscure sensitive information (Agrawal & Srikant, 2000; Clifton, Kantarcioglu, Vaidya, Lin, & Zhu, 2002; Dwork, 2008; Hill et al., 2013). The resulting dataset, sometimes referred to as a "limited dataset", can be shared more broadly. This form of sharing, called "statistical disclosure control" (Foster, 2018), seeks to allow research on data without ever obtaining access to information about individuals. For instance, vehicle in-cabin video may be transformed to obscure occupant faces while leaving their motions and actions visible. More refined options may capture facial expressions of the driver (e.g., road rage) while hiding his/her identity.

The third and final form, the *enclave*, is a controlled compute environment that enables the use of computational tools to analyze data while controlling both access and outputs. Vetted or known individuals are given access to the restricted data for the purposes of computational analysis, while the outputs of that computational analysis are strictly controlled. Foster (2018) differentiated between air-gapped enclaves from secure remote access enclaves where the former is disconnected from the Internet. In the latter, "the analyst connects remotely, for example over a virtual private network, to the data enclave." This paper focuses on the secure remote access enclave, which we posit is a complex socio-technical infrastructure. This paper puts forth the social and technical components of a remote enclave on equal footing and shows how the two trade off against each other.

All three options of Figure 1 (partial, transform, and enclave), limit either the information available or its access, that is, regulated are the *who*, the *what*, or the *where* or some combination of all the three. The "transform" option regulates the "what": accessed is limited to data that have been transformed in some way. The "partial" option limits access to portions of the data or metadata, thus also regulating the "what". The enclave, the third option, regulates both the "where" and the "who" by limiting both where the data can go and who has access.

This paper puts forth the capsule framework. As an instance of a remote enclave, the capsule framework is a socio-technical system (STS) with requirements that span hardware, software, human interaction, and policy. The framework supports analytical access to a massive, digitized collection of texts or volumes from research libraries. Analytical access cannot rely on ad hoc, informal arrangements between parties who share data only after having built trust through personal interaction. Our experience leads us to posit that all such remote

environments for accessing data – remote enclaves, data commons, etc. – are socio-technical systems.

The capsule framework protects the data from unintended uses or uses prohibited by law, policy, or licensing agreement. It is composed of policy and technology components working hand in hand to enable controlled access to a massive text corpus that is restricted by copyright. Our experience is synthesized in this paper as the inevitable trade-offs that must be made between security, trust, and usability.

The term "safe open science" captured in the title refers to safety as it applies to the data. The safety of a system is the application of engineering and management principles to achieve an acceptable level of risk for a system (Storey, 1996). Safe open science can then be seen as the application of these principles to achieving acceptable risk in opening data to limited forms of use.

An enclave for text analysis of large scale, restricted data is analogous to a playpen for toddler play. The playpen provides a safe environment; it does not restrict the type of toys that are in the playpen, size excepted, but it does control the comings and goings, keeping pets out and toddlers in. The analogy only holds so far, as toddlers are prone to toss toys out of the playpen, but the image of an environment providing protection is useful here. Like the playpen, the capsule uses the structure of the framework (the playpen walls) to allow access to the full data. The capsule controls what can be removed; the restricted data cannot be tossed out of the capsule. The compute resources of a capsule frequently reside close to the restricted data.

The capsule framework is a socio-technical system involving a controlled interaction between humans, machines, and the environmental aspects of the work system. Some of the interaction is captured through a *threat model* that captures the trade-offs that are made during the design of a system. The policies that are needed are influenced by the situation of use, which includes the restrictions on the data, assumptions of use, and the limits of the software services themselves. In addition, processes are required to enforce human activity compliance with policies.

This paper is based on our nearly decade-long experience in developing, deploying, and maturing a capsule framework implementation for computational analysis of a massive collection of digitized volumes (books and other materials) from academic and research libraries. The digitized texts are restricted by the in-copyright status of the volumes and by use restrictions on the metadata. Typically, around 38% of the digitized volumes are in the public domain, leaving over 60% of the content subject to legal restrictions on access.

In this paper, we take the reader through the social and technical aspects of the capsule framework developed for use in HathiTrust[2]. We discuss how we arrived at the policies that support the system and the tensions between security, trust, and usability. Any remote enclave is a set of policies, processes, software, and hardware; the particular manifestation each takes is a function of need. HathiTrust needs are dominated by three factors: i) the size of the collection, ii) the unknowable need for the kinds of tools used to analyze a massive textual corpus, and iii) the directive that any activity falls within *nonconsumptive research*.

We posit that the remote enclave is a viable third option for open science, that is, for making data "open as possible, as closed as necessary". We illustrate its use in HathiTrust through example uses. We further venture the uses of the capsule framework within the larger collection's management responsibility of academic and research libraries. In the remainder of the paper, we use the terminology of Foster (2018) to refer to the person using restricted data in a research project as an "analyst"; a term preferred to "user", "researcher", or "scholar", all of which are either demeaning or have ambiguous roles. This paper does not address the manner in which patient permission is secured for secondary use of data collected about patients.

2 Motivation

HathiTrust is a partnership of academic and research institutions, offering a collection of millions of titles digitized from libraries around the world. In 2011, the HathiTrust Research Center (HTRC) was formed in close partnership with HathiTrust to facilitate nonprofit and educational uses of the HT collection by enabling computational analysis of works in the public domain and on limited terms to in-copyright works from its collection. The digitized texts of HathiTrust which are in copyright (over 60% of the digital library, mostly post 1923) are considered sensitive in a way similar to how microdata can be considered sensitive, that is, needing protection from unwarranted disclosure.

HathiTrust makes available features extracted for the nearly 16 million books and other volumes under its stewardship. Features are notable or informative characteristics of a text, including part-of-speech tagged token counts, header and footer identification, and

2 www.hathitrust.org

various line-level information[3]. This is an example of a transformed dataset.

The capsule framework design for use on the HathiTrust collection began in 2011 with an award by the Alfred P. Sloan Foundation to pilot the framework for the express purpose of *nonconsumptive computational analysis* of the HathiTrust collection. Nonconsumptive research is discussed in more detail later. The final report of the Sloan-funded project is given in the study by Plale, Prakash, & McDonald (2015).

The capsule framework is driven by the requirement of an unknowable set of tools for analysis. That is, there is a preponderance of computational content-mining tools; the text analysis portal TAPoR, for instance, lists 913 text-mining tools (Text Analysis Portal, 2019). Given the impossibility of predicting in advance the tools an analyst may employ in working with the HathiTrust collection, the capsule project took the approach of not deciding for the analyst in advance, but instead, give trusted analysts a computer that resides within the environment of the HathiTrust, and let the analyst install and run their own text and data analysis tools (Zeng, Ruan, Crowell, Prakash, & Plale, 2014).

Remote data enclaves take various forms depending on the objective. We discuss a couple here to help illuminate the unique design choices of the capsule framework. The capsule framework itself derives its original security architecture and name from *storage capsules* (Borders, Vander Weele, Lau, & Prakash, 2009) of Atul Prakash at the University of Michigan. Additional examples include the following: Bose, 2013; Lane & Shipp, 2008; Stiles, Church, Smith, & Elings, 2014.

The Federal Statistics Research Data Center (RDC)[4] is a remote secure facility located at a research institution and managed by Census Bureau personnel. As per Federal SRDCs (2019), all of the data accessed at an RDC are physically located at the Census Bureau's main data center in Bowie, MD. Access to the data is provided to researchers at the RDCs via a thin client that can only display information from the server and accept mouse and keyboard input from the researcher. The data itself do not leave the Bowie data center. The thin client has no ability to download data from the server.

The Inter-University Consortium for Political and Social Research (ICPSR) makes restricted-use data available through a physical enclave, a Virtual Data Enclave (VDE), and an approved local researcher's computing environment (Mathur, Bleckman, & Lyle, 2017). A VDE allows the launching of a virtual machine on the analyst's local computer, but the software and data files remain on ICPSR's server.

The technical architecture of the capsule framework is conceptually simple as shown in Figure 2. An analyst is given access to a computer that is either a virtual machine (VM) or a lighter weight VM called a "container". This computer, called a "capsule", is located remotely from the analyst; the analyst accesses their capsule through tools such as a Virtual Network Client (VNC) that makes the remote computer appear as a desktop (window) on the analyst's own computer. The restricted data enter through an entry point, and the results exit through an exit point. A scholar accesses their capsule remotely. Tools reside in their capsule, either preselected or installed by the analyst themselves, or some combination of both.

2.1 Framework of the Capsule

The capsule framework is implemented by means of policy, processes, and software services. We discuss the requirements for computational access to restricted data that exist within HathiTrust and the resulting policy and infrastructure implementation that realizes the capsule framework. The objective of the paper is to illustrate through a real use case, the policy and implementation trade-offs that have to be made in implementing the capsule framework. The solution is frequently a synergy, or even a compromise, between software implementation and policy implementation.

2.2 Policy

The capsule framework as implemented for HathiTrust is influenced by a couple high-level requirements.

- Acceptable research is restricted to computational analysis that is performed on one or more volumes (digitized texts); unacceptable research, on the other hand, is research in which a human being reads or displays substantial portions of an in-copyright or rights-restricted volume to understand expressive content presented within the volume. This is "nonconsumptive research."
- The capsule framework must allow analysts to use their own tools. There is no common agreement

3 https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset
4 The Federal Statistical RDCs are "partnerships between federal statistical agencies and leading research institutions. They are secure facilities managed by the Census Bureau to provide secure access to a range of restricted-use microdata for statistical purposes only."

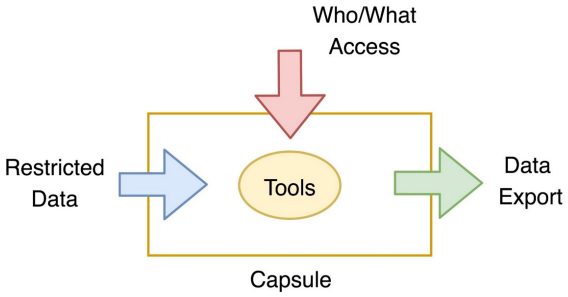

Figure 2. Sketch of an analyst's remotely located computer.

among interested researchers on any small set of analysis tools that would work for all analysts.

The first of these requirements, coined "nonconsumptive research", originates in the Authors Guild, Inc. and Association of American Publishers, Inc. et al. vs Google Inc. Amended Settlement Agreement[5] filed with the U.S. District Court Southern District of New York in 2009. The Amended Settlement Agreement ultimately failed, but the notion of nonconsumptive research was continued. The term "nonconsumptive research" is defined in section 1.93 of the Settlement Agreement. The second requirement derives from the sheer magnitude of the collection – now over 16 million volumes – and the completely new opportunity that computational access to it has provided.

We discuss each of the policies that are in place in HathiTrust to implement the capsule framework in the context of the unique requirements of HathiTrust[6].

Nonconsumptive Research Use Policy. *Nonconsumptive Research* is research in which computational analysis is performed on one or more volumes but not research in which researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand expressive content presented within that volume.

Nonconsumptive research includes image analysis, text extraction, textual analysis and information extraction, linguistic analysis, automated translation, and indexing and search, that is, nonconsumptive research is computational analysis that enables research uses of the volumes in HathiTrust, while protecting them from unwarranted disclosure, complying with the terms of copyright as well as the terms of the cooperative agreement that the research libraries have with Google for the content that Google digitized[7]. The full policy statement is available at https://www.hathitrust.org/htrc_ncup.

Use Agreement. A Use Agreement[8] exists between HT and individuals intending to use the capsule service. We call out three items as most illuminating of the synergies between policy and software that are highlighted in this paper:

- First, analysts agree to read and comply with the Nonconsumptive Use Research Policy and use their capsule for nonconsumptive research purposes. This *"no eyeballs on texts" requirement cannot be implemented completely in software without unduly restricting analyst activities in their capsule, so HT chose to strengthen adherence by means of restricting who has access to a capsule, verifying identity, and a strong use agreement.*
- Second, an analyst submits a form indicating their intended use and expected forms of outputs. *This clause ensures efficient and thorough manual review of all products (data) exported from a capsule.*
- Finally, by using the capsule framework, an analyst acknowledges that information about their activities while active within their capsule may be reviewed in manner consistent with HathiTrust privacy policy. *This clause informs the analyst that auditing is taking place, part of a "trust, but verify" philosophy.*

Rights database. A rights database stores and tracks rights information about each digitized volume in HathiTrust. At the core of the database is an algorithm that assigns rights to a volume. The algorithm considers a) copyright status and/or explicit access controls associated with the volume, b) volume's digitizing agent (e.g., Google or the University of Chicago), and c) identity of user (if known) in order to determine access rights. The rights database is used, for instance, to provide an analyst with a demo capsule that uses only public domain content. The policy can be found at https://www.hathitrust.org/rights_database.

Export review. A nonconsumptive export is a data product emerging as an output of computational analysis that meets the criteria of nonconsumptive research, that is, the exported data would pass results review. Nonconsumptive exports are released from a capsule through a specific action by the capsule analyst. Nonconsumptive exports from a capsule must be in

5 Amended Google Settlement, at the Internet Archive Wayback Machine (web.archive.org) by date of Dec 8, 2011 and URL of http://www.googlebooksettlement.com/Amended-Settlement-Agreement.zip

6 Not addressed in this study are the legal arrangements between the organizations (HathiTrust and HathiTrust Research Center) and their respective universities (University of Michigan and Indiana University) for hosting in-copyright content.

7 https://www.hathitrust.org/datasets

8 https://www.hathitrust.org/htrc_dc_tou

human-readable form (such as a Unicode or ASCII text or csv file). These exports undergo manual or automated review by HT staff prior to release in order to affirm its compliance with policy for nonconsumptive research.

3 Infrastructure

The technical architecture of the capsule framework in HathiTrust is a set of software services (Plale, Prakash, & McDonald, 2015), (Zeng et al., 2014) that collectively enable analysts to engage with restricted content through a capsule that is made available to them for a period of time. HathiTrust allows a combination of canned and analyst-installed tools; the latter can be installed by a researcher in their capsule without prior vetting of the tool. Through the vehicle of a use agreement, obtained from the analyst is a description of their anticipated tool use for subsequent review of exported results.

It is the rare application that analyzes more than a few 1000 digitized texts, or a million texts at the most. These texts are carefully chosen prior to analysis, and they are built up into what HathiTrust calls a Workset (Jett, Cole, Maden, & Downie, 2016). A workset is a list of volumes annotated with semantically rich metadata with which an analyst is working.

The capsule framework runs on physical servers located at Indiana University Bloomington (IUB) and accesses restricted data that reside in IUB highly secure Data Center.

3.1 Threat Model

Data protection exhibits the classical trade-offs found in securing software systems (Hasan, Myagmar, Lee, & Yurcik, 2005). For instance, the use of encryption may provide confidentiality, however, it may also hamper performance and usability. A threat model is a high-level articulation of the trade-offs that are made during the design of a software system to capture the system guarantees. A threat model for HathiTrust is provided here, although it should be noted that the threat model is not an implementation guide and the actual security mechanisms used in HathiTrust Research Center are documented in a detailed, internal security implementation document that is vetted by experts at the level of the chief security officer of the institution.

The threat model for the capsule framework implementation in HathiTrust is built on the assumed existence of a Trusted Computing Base (TCB), wherein the totality of security mechanisms within a secure system resides (Rushby, 1984). A common means to access a computer remotely is through a Virtual Network Computing (VNC), which gives an analyst desktop access to a remote machine, that is, the remote machine is accessible through a window on the analyst's own computer.

The threat model of the capsule framework as implemented in HathiTrust can be captured by the following eight statements:

A. An analyst accesses restricted data using his or her assigned virtual machine.
B. The analyst's assigned capsule is not trusted. Other resources that support the analyst's capsule are trusted, including the Virtual Machine Manager (VMM), the host that the VMM runs on, the system services that enforce network and data access policies for the virtual machines, and the data storage system. All are within the Trusted Computing Base.
C. We assume the possibility of malware (i.e., malicious software) being installed as well as other remotely initiated attacks on the VM. These attacks could potentially compromise the entire operating system and install a rootkit, both of which are undetectable to the analyst.
D. Analysts are themselves considered to act in good faith, but this does not preclude the possibility of them unwittingly allowing the system to be compromised. This is a reasonable assumption and motivates why analysts are required to sign a use agreement before using the system.
E. Analysts are working within their assigned capsule work within two modes: secure mode and maintenance mode. When in *maintenance mode*, an analyst has complete freedom to upload and download material from the Internet but cannot access the protected data. In *secure mode*, the analyst is given direct access to the restricted materials but cannot access the Internet and is prohibited from downloading or copying from their capsule to their desktop.
F. While in maintenance mode, analysts have a graphical interface to the machine. They also have remote command-line access (i.e., SSH access) in order to upload datasets and install software more easily. However, command-line access does provide a channel for potential data leak which is addressed through employment of a signed use agreement and establishment of a profile for each analyst.
G. A potential threat is that of covert channels between virtual machines that run on the same host machine. A solution requires using two physically separated systems: one that only runs capsules that are in

secure mode and a second that runs capsules that are in maintenance mode. HT currently performs routine host port scanning.

H. The analyst's state is retained in a capsule across sessions of work, but when an analyst completes his/her work and wants to pull data out of the capsule, he/she must store the results to a special drive. The contents of this drive are manually reviewed before results are made available to the analyst.

3.2 Capsule Architectural Details

The capsule framework as it is implemented in HathiTrust consists of three architectural layers as shown in Figure 3. The layers are referred to, from bottom to top, as a back-end layer, a web service, and a web front end. We describe the functioning of each of these abstract layers but do so in the context of high-level descriptions of the software packages that implement the abstract layers. The back-end layer is packaged into a *Hosts package* and *Image package*. The web service later (middle of figure) is packaged as the *Web Services package* and the top layer as the *Web UI package*.

Host Package. The host package consists of all the software needed to manage capsules for a single compute server (an implementation of the capsule framework will utilize numerous compute servers simultaneously to support a larger number of analysts.) This includes a virtual machine manager that manages the individual capsules and the quantity of resources each is allowed to use (e.g., cores, memory, and network). Firewalls on the compute server must be precisely configured to ensure data security and prevent data leakage.

The restricted data collection resides on a separate back-end storage server. It is accessed by means of scripts inside an analyst's capsule. An analyst's identity is passed to the back-end storage server so that auditing tools can detect and trace malicious activity. HathiTrust additionally carries out regular scans of capsule server activity to detect vulnerability and verify the ongoing accessibility of individual capsules.

Image Package. Several software packages are preinstalled and settings pre-configured in a capsule to allow it to communicate with the compute server on which it is running. An automation tool is available to generate a pre-built capsule image. HathiTrust uses a tool called Packer[9] to build images for analysts in advance. Packer is also used to pre-load a public domain subset of texts for use in the classroom.

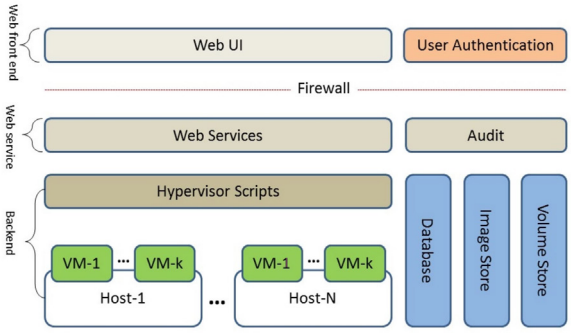

Figure 3. Architecture supports k∗N individual VMs running in the back-end layer and managed by a virtual machine hypervisor.

Web Services Package. This package consists of a Web Service interface and management databases. It maintains information about the status of capsules (e.g., to whom they are checked out) in its database. HT employs Docker[10] to host both the Web Service and the relational (i.e., mySQL) database in separate containers. Other container approaches would work equally as well.

Web UI Package. The Web User Interface (UI) consists of two dashboards: a DC Dashboard that is used by an analyst to manage their own capsules and a Reviewer Dashboard. The DC Dashboard supports capsule creation, starting, stopping and mode switch (between maintenance mode and secure mode). It also previews the individual capsule information such as resource allocations, host, connection information, etc.

The Reviewer Dashboard is used by HathiTrust administrative staff to review the data that an analyst wishes to export from their capsule upon completion of analysis. Having an administrative staff member responsible for the review of results ensures that the restricted data are not inadvertently released at the completion of a computational analysis activity.

4 Use in HathiTrust

The capsule framework in HathiTrust provides for controlled access to over 16,000,000 volumes of content available in the HathiTrust Digital Library. As of this writing, 160 analysts are working with 181 capsules. 209 capsules were created in the first six months of 2018,

9 Packer automation tool https://www.packer.io/

10 Docker https://www.docker.com/

some of which have since been deleted by analysts upon completion of their work. The latest round of awardees in the HTRC Advanced Collaborative Support program (HTRC ACS Awards, 2017) provides an illustrative snapshot of the sorts of research questions being explored through HathiTrust collections in the capsule environment. Seven individual analysts or analyst teams are exploring topics such as community reading programs, the history of book design, U.S. women's movements, history of the U.S. novel, literary novelty, the Iowa Writers' Workshop, and enhancement of metadata for the Oxford English Dictionary (HTRC ACS Awards, 2017).

The Iowa Writers' Workshop project is a particularly interesting use case. Analysts have assembled a corpus of works by authors affiliated with the renowned University of Iowa Writers' Workshop. The analyst team is performing analysis on that corpus to determine whether a "Workshop style" exists and what the characteristics of such a style might be (White, 2018). The analysis measures formal features and collects metrics such as "vocabulary size, sentence length, or even frequency of male and female pronouns" (Kelly, 2017). They also track location references in the texts, "making it possible for the project to analyze regional representation trends in literary works" (Kelly, 2017).

The subjects and disciplines represented by the ACS awardees are largely from the domains of language and literature, with the analysts coming from departments in the humanities or, at least in one case, the social sciences. In many cases, these humanities or social science scholars have sophisticated technical skills. Moreover, analyst teams may be interdisciplinary and include faculty from computer science or information science or librarians with expertise in digital scholarship and digital humanities methods.

However, because so many analysts working with the HathiTrust collections are from humanities disciplines that have not traditionally adopted computational research and analytical methods, HTRC works to provide various points of entry for research with a capsule. For instance, within the capsule, HTRC has successfully implemented Voyant Tools "a web-based text reading and analysis environment designed to facilitate reading and interpretive practices for digital humanities students and scholars as well as for the general public" (Sinclair & Rockwell, 2016). Voyant tools run on a locally hosted Web server, within the secure environment of the capsule, allowing analysts to take advantage of the GUI interface and Web environment, as an alternative or complement to the command-line tools typically used for text analysis in the capsule.

5 Use in Library's Special Collections

Libraries have a long history of curating and making available special collections of books, letters, newspapers, and other materials. As these special collections become increasingly digital, how well suited is the capsule framework for access and management of digitized special collections? In a sister project[11], we partnered with several research libraries to study the current library needs and practices in provisioning services for computational access to special collections and to extend the capsule service to enable secure access to restricted data in libraries.

Enabling access to library collections as data is an emerging area of research and practice. The HathiTrust Research Center set an example with its early technical, policy, and security frameworks that defined nonconsumptive research and provided tools for analyzing digitized texts while restricting humans from reading or downloading the text. The Library of Congress transforms some of its collections into a machine-readable form and offers an online space that supports experimentation via APIs, bulk downloads, and human coding to help users work with large collections computationally (LC for Robots, 2018). A growing number of cultural heritage institutions are interested in transforming their collections into data as they pledge to support computationally driven research and teaching (Santa Barbara Statement, 2018) [12]. These projects provide insight into how making collections ready for computational analysis makes them more relevant and useful, for example, by broadening public access or studying the role of underrepresented groups in history.

Each project works with one or several libraries, engaging with unique collections and addressing their specific challenges. Our partners have identified several unique special collections that differ in their formats (e.g., video, news broadcast, and text) and types of restrictions (e.g., vendor licensing, copyright, or human subject confidentiality). More research into the practices of stewards and users is needed to identify common needs and develop best practices in policies and technologies. Our research and interactions with librarians at partner institutions and collaborative events point to the following common needs in developing technology to support computation on special collections:

11 Data Capsule Appliance for Research Analysis of Restricted and Sensitive Data in Academic Libraries, IMLS LG-71-17-0094-17, https://www.imls.gov/grants/awarded/lg-71-17-0094-17

12 https://collectionsasdata.github.io/part2whole/cfp/

A. **Flexible access controls enabled by both policy and technology.** Although libraries seek to maintain their collections as open as possible, they recognize legal, ethical, and other obligations that may prevent them from offering open access. For example, the collection may come with licensing requirements that limit a number of users, user affiliation, or types of use (e.g., educational). Otherwise, a donor may still be living and require special permission or restrict access altogether to parts of a collection, as in the case with the papers of Henry Kissinger maintained by the Yale library[13]. User categorization, content copying, online availability, and access to unprocessed materials all need to be updated in light of increasing digitization and computational use.
B. **Use and preservation.** Librarians and curators are interested in providing a long-term access to their collections; they often need to balance the needs of current use and discoverability with the requirements of preservation. Digital environments can serve both, but they need to become part of the collection management lifecycle and fit with the existing processing workflows. Although the growth of digital collections may be encouraging from the perspective of data analysis and computational techniques, sustainability, scalability, and equity of access over time have to be considered too.
C. **Metrics for usefulness and usability.** Libraries are increasingly focusing on addressing user needs and tracking uses of their services and collections. An OCLC survey of special collections (Dooley and Luce, 2010) pointed to the increasing uses of special collections and to the changes in patterns of use, including the uses of audio and video materials as well as the use of digital methods of access. On the one hand, digital environments can provide tools for tracking users, and on the other, the development of such environments needs to be connected to user communities and their interests. Without better understanding of who is interested in computational analysis of collections and for what purpose, a library risks a poor return on investment in its provisioning for computational analysis of the collection.

The process of adoption of a new technology includes the assessment of the fitness of the technology against the need (Rogers, 1962). Such an assessment is particularly important when technology is being considered by innovators and early adopters – two groups from the Rogers' model that are interested in trying out new technologies and are willing to take the risk. Acceptance of technology has also been shown to depend on meeting the expectations of performance and effort, on social influence, and on facilitating conditions, such as technical and organizational support and skill level of adopters (Venkatesh, Morris, Davis, & Davis, 2003). The fitness of the capsule framework against the needs of libraries and their special collections has been determined through partner consultations. Our partners are innovators and early adopters who are interested in evaluating the capsule framework through direct assessments of the software and "trying it out" as part of the libraries' workflows.

For a sizeable system such as the capsule, "trying it out" involves iterations of installation and system administration and close interactions between technical and library teams. Our case studies reveal that research libraries may not have computers or staff available to undertake such hands-on evaluation. A test bed that can be set up in a cloud environment and tested without affecting the operational systems will help reduce barriers to test and meet performance and effort expectations.

A new system also needs to be further evaluated for adoption by assessing the skills required, especially on the service provider side. Competence of skills in firewalls, Internet protocols, and the provision of administrative privileges that enable or disable security measures more frequently exist in IT unit of an academic organization rather than in the library. At the same time, they are archivists or librarians who must decide what levels of access to collections are needed, taking into consideration both users' need and restrictions imposed by collection owners and donors. *Therefore, implementation of an instance of the capsule framework within an academic library becomes a collaborative university effort, where access and use decisions depend on institutional policy and infrastructure.*

6 Conclusions

The capsule framework is a new approach to accessing and sharing restricted data that protect data from unintended use or uses prohibited by law, policy, or licensing agreement while allowing computational access to restricted data by known individuals and under controlled circumstances. The capsule framework in HathiTrust is used to enforce nonconsumptive research use and does so through policy, processes for manual review of results, and security built into the technical infrastructure. Security includes that of

13 https://web.library.yale.edu/digital-collections/kissinger-collection

the capsule framework itself, and the computers, storage servers, and network and software services that manage and move data in and out of analyst VMs.

The capsule framework is a viable solution to accessing restricted data, and in some senses the most promising of the three fundamental approaches: partial, transformed, and capsule. Other forms limit the "what" by obscuring the data in some way through differential privacy, anonymization, statistical aggregation, etc. or limit the "who" based on need to know (e.g., federal security levels). The capsule framework is a limit of "where" in that there's a virtual place where work needs to take place, and to some extent a "who" limit as well. The benefits of "where" security is in richer access to the data; the challenge, however, is in making the environment friendly (and fast) while enforcing the protections on the data.

For the capsule framework to become an accepted form of Open Science data sharing, there must be seamless awareness of the increasing forms of sharing starting with the Institutional Review Board (IRB) at academic institutions. An IRB study designed to protect data about individual human participants will generally dictate the ways in which data are to be transformed or kept hidden entirely. Data repositories that provide protected storage and data preservation must equally aware and support commonly accepted levels of access. The Dataverse repository (https://dataverse.org/) has begun this effort.

As demonstrated through the HathiTrust use case, the capsule framework is implemented through synergistic policies, processes, and technologies, encoded in a Threat Model. The decision of how a security check is implemented in policy versus in software is a function of a concerted effort to retain some levels of ease and efficiency of use for the user. It is an open question as to how generalizable is the set of software policy trade-offs adopted for nonconsumptive research for other types of restricted data. This open question is deserving of more study.

In working with our library partners, the issue of text-mining across collections from multiple publishing platforms arises. For example, what is an appropriate policy and technical interface to data when one collection resides in a file system and the other in an SQL database? Open source data virtualization layers (such as Teiid, http://teiid.io/) may provide an answer. In addition, what is a model for library management of such a secure enclave described here? Incorporating capsules and computational access to restricted collections into library-based services will need to fit within the whole library organizational structure to avoid having a separate point person who becomes "jack of all trades" and helps with technical, data mining, and information retrieval or reference questions.

The capsule framework as implemented for HathiTrust has dependencies that limit its scalability to, say, cloud resources. An open question is this: *"What is a new technology design that implements the capsule framework, and HathiTrust's specific threat model, without the scalability and portability limitation of the current system?"* There is a need in HathiTrust to extend the capsule framework to run on cloud platforms such as Jetstream, Amazon Web Services, or university cloud resources to give analysts access to more compute resources to analyze even larger portions of the HathiTrust collection at any one time.

The Data Capsule Host automation scripts are available at DC Host automation (2018) and the Data Capsule Image automation repository is available at DC Image automation (2018).

Acknowledgments: This research is supported in part through funding from the Institute of Museum and Library Services (IMLS) #LG-71-17-0094, the Andrew W. Mellon Foundation #41500672, HathiTrust Board of Governors, The Alfred P. Sloan Foundation #2011-06-27, and a Lilly Endowment grant to Pervasive Technology Institute at Indiana University.

We thank Atul Prakash, University of Michigan, for foundational contributions to the data capsule model. We thank HathiTrust's past and present colleagues: J. Stephen Downie, Mike Furlough, Beth Sandore Namachchivaya, Harriett Green, and John Unsworth. We thank the reviewers for the thoughtful feedback that has substantially improved this manuscript.

References

Advanced Collaborative Support (ACS) Awards. (2017). Retrieved from https://www.hathitrust.org/htrc_sp17acs_awards

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. [ACM.]. *SIGMOD Record, 29*(2), 439–450.

Borders, K., Vander Weele, E., Lau, B., & Prakash, A. (2009, August). Protecting confidential data on personal computers with Storage Capsules. In *Proc. of 18th USENIX Security Symposium*(pp. 367-382). CA, USA: USENIX.

Bose, R. (2013, May). Implementing a Secure Data Enclave with Columbia University Central Resources. In *IASSIST Conference*. Germany.

Bureau, U. S. (2019). Federal Statistical Research Data Centers (RDCs). Retrieved from https://www.census.gov/about/adrm/fsrdc/about/hostrdc.html

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. [New York, NY, USA: ACM.]. *SIGKDD Explorations, 4*(2), 28–34.

Dooley, J. M., & Luce, K. (2010). *Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives* (p. 153). Dublin, OH: OCLC Research.

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation, 4978, 1-19*. Berlin, Heidelberg: Springer.

Foster, I. (2018). Research infrastructure for the safe analysis of sensitive data. *The Annals of the American Academy of Political and Social Science, 675*(1), 102–120.

Hasan, R., Myagmar, S., Lee, A. J., & Yurcik, W. (2005, November). Toward a Threat Model for Storage Systems. *Proc. of 2005 ACM Workshop on Storage Security and Survivability* (pp. 94-102). New York, NY, USA: ACM.

Hill, R., Hansen, M., Janssen, E., Senders, S. A., Heiman, J. R., & Xiong, L. (2013, July). An empirical analysis of a differentially private social science dataset. In *Proc. of PETools: Workshop on Privacy Enhancing Tools, Held in Conjunction with the Symp on Privacy Enhancing Tools Symposium,* Bloomington.

Jett, J., Cole, T. W., Maden, C., & Downie, J. S. (2016). The HathiTrust Research Center workset ontology: A descriptive framework for non-consumptive research collections. *Journal of Open Humanities Data, 2*, 1–7.

Kelly, N. M. (2017). Text Analysis Tool Samples. Retrieved from https://newreadia.wordpress.com/text-analysis-tool-samples/

Lane, J., & Shipp, S. (2008). Using a remote access data enclave for data dissemination. *International Journal of Digital Curation, 2*(1), 128-134.

LC for robots. (2018). Retrieved from Library of Congress Labs: https://labs.loc.gov/lc-for-robots/

Mathur, A., Bleckman, J. D., & Lyle, J. (2017). Reuse of Restricted-Use Research Data. In L. Johnston (Ed.), *Curating Research Data, Volume Two: A Handbook of Current Practice* (pp. 258–261). Chicago: Association of College and Research Libraries.

Plale, B., Prakash, A., & McDonald, R. (2015). *The Data Capsule for Non-Consumptive Research: Final Report.* Tech. rep., Indiana University Bloomington.

Rabesandratana, T. (2013, November 21). European Parliament Approves Horizon 2020 Funding Plan. *Science Magazine*.

Rogers, E. M. (1962). *Diffusion of innovations (1st ed.)*. Glencoe, New York: Free Press.

Rushby, J. (1984). A Trusted computing base for embedded systems. In *Proc. of 7th Annual Dept of Defense/NBS Computer Security Conference* (pp. 294-311). Gaithersburg, Maryland.

Sinclair, S., & Rockwell, G. (2016). Voyant Tools. Retrieved from http://docs.voyant-tools.org

Stiles, J., Church, J., Smith, E., & Elings, M. (2014). *UC Berkeley Library: Report of the Restricted Use Data Task Force*. Retrieved from http://www.lib.berkeley.edu/AboutLibrary/reports/Restricted-Use-Data-Task Force-Report0914.pdf.

Storey, N. R. (1996). *Safety Critical Computer Systems*. Boston, MA, USA: Addison Wesley Longman Publishing Co., Inc.

TAPoR - Text Analysis Portal for Research. (2019). Retrieved from http://www.tapor.ca/

The Santa Barbara Statement on Collections as Data. (2018). Retrieved from https://collectionsasdata.github.io/statement/

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly, 27*(3), 425–478.

White, N. (2018). *Breaking down the HTRC Data Capsule*. Retrieved from https://dsps.lib.uiowa.edu/programera/2018/02/09/breaking-down-the-htrc- data-capsule/

Withana, S., Plale, B., & Kouper, I. (2018). *Data Capsule Host automation scripts*. Retrieved from https://github.com/Data-to-Insight-Center/Data-Capsule- Appliance-Host/

Withana, S., Plale, B., & Kouper, I. (2018). *Data Capsule Image Automation*. Retrieved from https://github.com/Data-to-Insight-Center/Data-Capsule-Appliance-Guest/

Zeng, J., Ruan, G., Crowell, A., Prakash, A., & Plale, B. (2014, June). Cloud computing Data Capsules for non-consumptive use of texts. In *Proc. 5th ACM Workshop on Scientific Cloud Computing* (pp. 9-16). New York, NY, USA: ACM.