
# Secure Text Analysis at Scale Over Sensitive Text : HTRC Data Capsules

Use case: RDA Digital Humanities Workshop, May 2015

While the exact percentage changes daily, nearly 70% of the digitized texts in HathiTrust are under copyright. HathiTrust approved the creation of the HathiTrust Research Center in 2011 to provision computational investigation (e.g., text mining) over the full corpus of copyrighted and non-copyrighted content. The challenges of provisioning a research commons to support secure analytics over objects in the HT digital library is the subject of this use case. A companion case gives a brief overview of the mainstream tools and services of HTRC as they currently exist in HTRC's "SHARC", Secure HathiTrust Analytics Research Commons (SHARC)[1].

**Starting principles**

We take for granted that computational analysis (text mining) of a large digital repository of over 13 million volumes is useful for research/scholarship, education, and new forms of exploration of content. Were this not the case, HTRC would not be interesting to a sustained community of about 100 users who regularly attend its yearly HTRC UnCamp and participate in its user group meetings. Our experience suggests an 80/20 rule applies in that 80% of the use cases of computational analysis of HathiTrust Digital Repository are over content the size of a person's personal readings (1200 books) and the remaining 20% at larger scale. Experience further suggests that the 80% are dominantly in the digital humanities and 20% are informatics kinds of uses, but the line between DH and informatics is blurred so the rule has many exceptions.

In provisioning computational analysis over HT, philosophy of HTRC is to meet the needs of the 80% by giving researchers tools for independent investigation, and working hand in hand with the remaining 20% to accomplish their goal (the heroic cases). This may change over time.

The digital nature of the content and restrictions of copyright precludes a model whereby a researcher checks out a set of (digital) volumes/books, takes them home (with promising to return them in morning). Can HTRC provision dedicated compute resources in proximity to the data so that computational analysis can take place in a relatively free form fashion, while protections on the data actively work to ensure data are not leaked, or if leaked, have no value?

**HTRC Data Capsule concept**

The Data Capsule concept is to provide access to restricted datasets where it is desirable to provide access to the datasets to end-users who are generally trusted with remote access to the dataset. End-users can bring in their own software for analysis as well as bring in external datasets. The security risk that the system mitigates is that of software that is used for analysis being inadvertently malicious and thus leaking data to a third party. Our current prototype permits remote access to the datasets. In situations where datasets are highly sensitive, the proposed technology could be used in more restricted

[1] http://www.hathitrust.org/htrc

settings where end-users can access the data only from secure rooms, but still have the ability to import their software and external datasets to assist in the analysis over the network from the secure facility, without introducing the risk of leaking restricted datasets.

**How it Works**

- A researcher "checks out" a virtual machine (VM) that is pooled in the SHARC environment and is loaded with familiar tools for analysis
- The VM runs in, and only in, the HTRC SHARC environment
- A researcher owns their VM through weeks/months of analysis
- A researcher is able to install their own tools in addition to what comes preconfigured.
- Core to the capsule concept is its dual mode operation. That is, it has a Secure mode and an Open mode; security turns on and off when the mode is switched from Open to Secure and back again. During install, the VM is in Open mode, so open to the Internet for downloads/uploads, to FTP and SSH, but closed to the raw HathiTrust data. When analysis is about to begin, the researcher enters a command and the capsule switches modes from Open to Secure. In Secure mode, the APIs to the raw data become operational. SSH is still available, but is limited to a connection to the researcher's machine in his remote location.
- Results of analysis in a Capsule are stored to a particular mounted directory that lives through the transition from Secure back to Open mode. Results are manually reviewed before being mailed to the researcher.
- Getting data and tools into a VM is easy, but there is a controlled and audited process for getting results out of the VM

**Challenges**

Data Capsules is HTRC's vehicle through which researchers will get direct access to the raw copyright content. It is operational in beta form in HTRC SHARC V3.1. Further improvements to the system are several:

a) Support a stronger threat model so that the user does not have to be trusted,
b) Create pre-configured VMs for a particular domain (e.g., computational linguists share a set of common tools; statistical analysis is best done in R; feature extraction needs a noSQL server instance installed etc.)
c) In SHARC v3.1 a Data Capsule researcher is limited to working within a single VM and VMs do not communicate. This limits the compute power they can potentially take advantage of.
d) Further protections on the data are helpful; protections that actively work to ensure data are not leaked, or if a leak occurs, the data have no value.

With thanks to the Alfred P. Sloan Foundation

Use Case Table

| | |
|---|---|
| ID: | |
| Title: | Secure Text Analysis at Scale Over Sensitive Text : HTRC Data Capsule |
| Description: | Secure environment for computational analysis on HathiTrust volumes |
| Trigger: | - |
| Preconditions: | The 13M+ volumes in HathiTrust digital library are a rich wealth of unexplored (computationally) content. Yet about 70% of it is under copyright, requiring that the data be treated as sensitive. |
| Steps for Main Success Scenario: | The success scenario for the HTRC Data Capsule is a security and privacy enabled research commons environment in which a researcher can carry out text analysis using their or their community's tools and interact with the texts as closely as they need to while at the same time there are mechanisms operating under the covers that ensure that the protections stated in the Capsule threat model are not violated. |
| Alternate scenarios: | - |
| Postconditions: | May need to read full text for interpreting results |
| Frequency of Use: | The analysis can be done as frequent as needed |

| | |
|---|---|
| Status: | Draft |
| Author: | Beth Plale, Miao Chen |

References

HathiTrust Research Center. http://www.hathitrust.org/htrc

Zeng, J., Ruan, G., Crowell, A., Prakash, A., & Plale, B. (2014, June). Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing* (pp. 9-16). ACM.