Data Capsule Appliance for Research Analysis of Restricted and Sensitive Data in Academic Libraries

December 11, 2017 I CNI Fall Members Meeting

Robert H. McDonald, Indiana University
Erik Mitchell, University of California, Berkeley
John Unsworth, University of Virginia
Inna Kouper, Indiana University

DATA TO INSIGHT CENTER
PERVASIVE TECHNOLOGY INSTITUTE

LIBRARIES

BERKELEY Library UNIVERSITY OF CALIFORNIA

UNIVERSITY of VIRGINIA LIBRARY

HATHI TRUST Research Center

HATHI TRUST RESEARCH CENTER

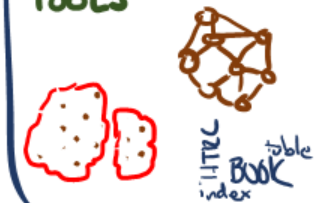
EXTRACTED FEATURES

PARTS OF SPEECH, WORD COUNTS, ...

?? WHICH IS BEST CHOICE FOR MY NEEDS ??

LITRC PORTAL

CANNED ANALYSIS TOOLS

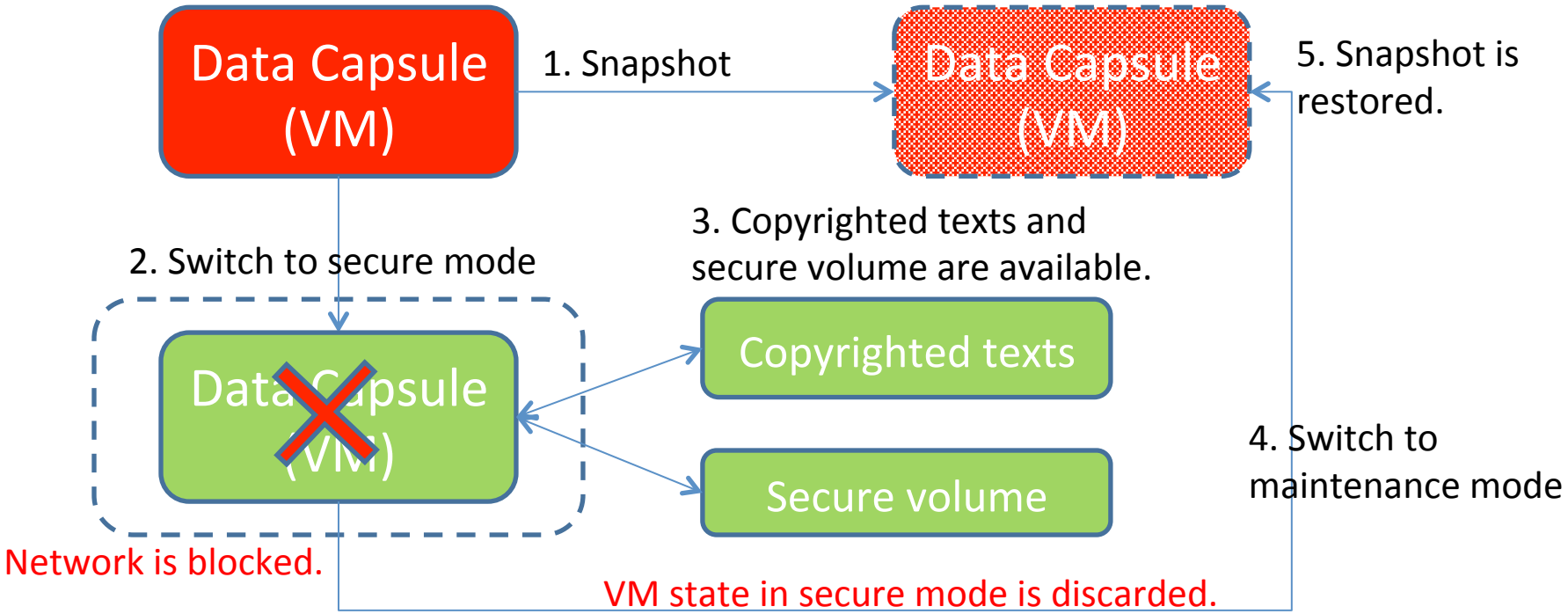


DATA CAPSULE

MY ANALYSIS TOOLS

HATHI TRUST DIGITAL LIBRARY (HTDL)

Data Capsule Mode Switch



Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing* (ScienceCloud '14). ACM, New York, NY, USA, 9-16. DOI=10.1145/2608029.2608031 http://doi.acm.org/10.1145/2608029.2608031

Photo by frank mckenna - Creative Commons No known copyright restrictions, https://unsplash.com/@frankiefoto?utm_source=ihaikudeck&utm_medium=referral&utm_campaign=wallpaper+feed

Our Project

Data Capsule Appliance For Research Analysis Of Restricted And Sensitive Data In Academic Libraries

Goals:

1. Understand library needs and practices in providing computational access to restricted collections
2. Extend an existing service to enable access to restricted data in libraries
3. Identify gaps in skills needed to enable secure data analytics

Activities:

- Design-oriented partner engagement
- Software architecture and evaluation

Project Deliverables

- Knowledge about restricted collections and their policies and contexts of use
- Data Capsule packaged as an appliance with support for new collection types and use cases
- Understanding of library-tech-research collaborations
- Emerging sense of community

UVA Use Case

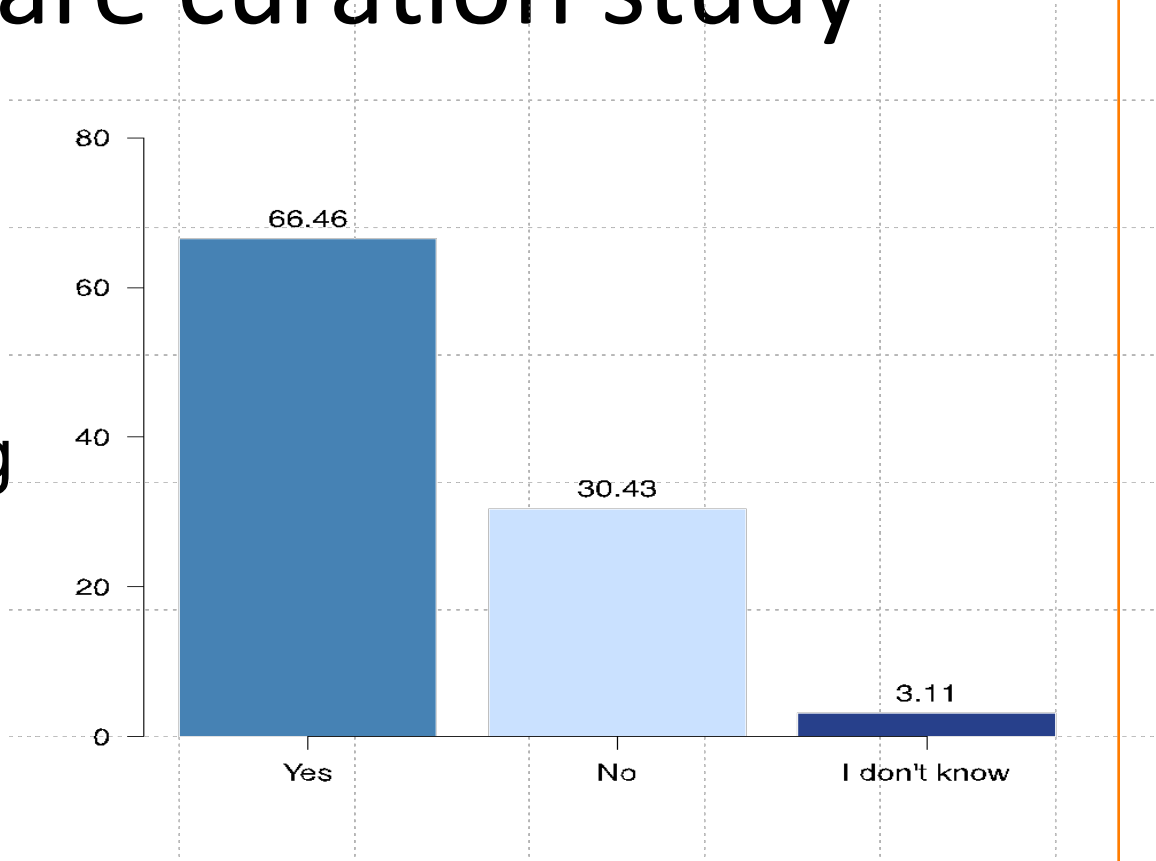
- Running captioning software on in-copyright video collections (with new urgency, now that Apple has bought the Pop Up Archive captioning service)
- Working with others in the project to create push-button installation of the Data Capsule software
- Understanding how something like the Data Capsule might factor in to other library projects, like …

UVA Context

- Work with the Library of Congress to provide computational access to remote collections of copyrighted digital media (held at the LOC's Packard campus)
- Work with Ithaka to understand the perceived demand and available support for text data mining, as seen by scholars, librarians, and publishers
- Work with Portico and JSTOR to explore text-mining across distributed collections of copyrighted material

Software curation study

The majority of the 30% who couldn't share related files indicated that this because of licensing issues or the data is sensitive.



AlNoamany, Borghi, Chassanoff, Thorton. Software as a well-formed Research object. Digital Library Forum. October 24, 2017.

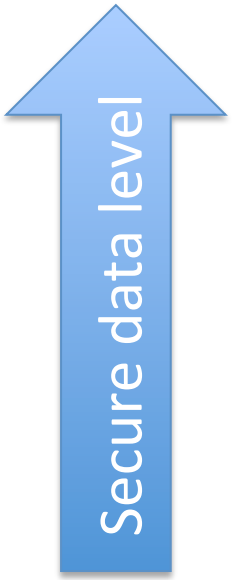
LIS data sharing study

Data ownership and privacy issues are barriers to data sharing

Mitchell, Moulainson-Sandy, Corrado. Making a case for Open Research. ASIS&T conference. October 2017

Analyzed subset made available	1
Data already available (e.g. public domain, in archive)	3
Data privacy issues	13
Desire to preserve ownership rights	1
Did not ask participants for permission	1
IRB issues	7
Lack of time	20
No incentive	1
Not common in our field	1

Where HTRC fits @ UCB



Technical environment	Use case
HTRC	**Packaged virtual environments including data protected from researchers**
”Cold room”	Non-connected compute environments
Analytic Environments on Demand	Virtually partitioned streamed desktop with connected secure storage
Savio	High performance computing

Technical Discussion and Models for Restricted Data

- Standard platforms and Tools
  - Machine independence
  - Push-button installation
- Security requirements evolve faster than other parts of the problem
- Data mining across collections without co-location solves policy problems

Discussion Topics

- Dealing with sustainability for the software and infrastructure: few examples of research projects that produced lasting library infrastructure (ArchivesSpace, DuraSpace, MONK?)
- More needs to be known about the use cases for libraries and archives
- Does the Data Capsule offer any way of dealing with new e-licensing platforms that allow use of copyrighted data only in closed environments?

DC Appliance Acknowledgements

DCA@IU

- Inna Kouper
- Robert H. McDonald
- Sachith Withana
- Marie Ma

DCA@UVA

- John Unsworth
- David Goldstein

DCA@UCB

- Erik Mitchell
- Yasmin AlNoamany

DCA Tier 2 Partner Libraries

- Mary Mellon, Indiana University
- Anne Houston, Lafayette College
- Kari Smith, MIT
- Francesca Gianetti, Rutgers University
- Nabil Kashyap, Swarthmore College
- Peter Broadwell, UCLA

Funding Agency

INSTITUTE of Museum and Library SERVICES

IMLS Grant
#LG-71-17-0094-17

DATA TO INSIGHT CENTER
PERVASIVE TECHNOLOGY INSTITUTE

LIBRARIES

BERKELEY Library
UNIVERSITY OF CALIFORNIA

UNIVERSITY of VIRGINIA LIBRARY

HATHI TRUST
Research Center

HTRC Acknowledgements

HTRC @ Indiana:

- John Walsh-Co-PI
- Beth Plale, Sr Science Advisor
- Robert McDonald
- Marie Ma
- Samitha Liyanage
- Leena Unnikrishn
- Jaimie Murdock
- Zong Peng
- Angela Courtney
- Leanne Nay

HTRC @ Illinois:

- J. Stephen Downie-Co-PI
- Harriett Green
- Tim Cole
- Jacob Jett
- Boris Capitanu
- Eleanor Dickson
- Ryan Dubnicek

Liz Lorang
University of Nebraska

Leen-Kiat Soh
University of Nebraska

David Mimno
Cornell University

HTRC UnCamp 2018, UC Berkeley
Jan 25-26, 2018
https://goo.gl/6oufPL