Towards Publishing Secure Capsule-based Analysis

Short Paper

Jaimie Murdock
Indiana University
School of Informatics and Computing
jammurdo@indiana.edu

Jacob Jett
University of Illinois
School of Information Sciences
jjett2@illinois.edu

Tim Cole
University of Illinois
School of Information Sciences
t-cole3@illinois.edu

Yu Ma
Indiana University
School of Informatics and Computing
yuma@iu.edu

J. Stephen Downie
University of Illinois
School of Information Sciences
jdownie@illinois.edu

Beth Plale
Indiana University
School of Informatics and Computing
plale@indiana.edu

ABSTRACT

Computational engagement with the HathiTrust Digital Library (HTDL) is confounded by the in-copyright status and licensing restrictions on the majority of the content. Because of these limitations, computational analysis on the HTDL must either be carried out in a secure environment or on derivative datasets. The HathiTrust Research Center (HTRC) Data Capsule service provides researchers with a secure environment through which they invoke tools that create, analyze, and export non-consumptive datasets. These derivative datasets, so long as they do not reproduce the full-text of the original work, are a transformative work protected by Fair Use provisions of United States Copyright Law, and can be published for reuse by other researchers, as the HTRC Extracted Features Dataset has been. Secure environments and derivative datasets enable researchers to engage with restricted data from focused studies of a few dozen volumes to large-scale experiments on millions of volumes. This paper describes advances in the Capsule service through a case study of how the HTRC Data Capsule service has advanced our activities on provenance, workflows, worksets, and non-consumptive exports through a topic modeling example. We also discuss the potential applications of this Capsule-based model to other digital libraries wrestling with research access and copyright restrictions.

CCS CONCEPTS

•Information systems → Digital libraries and archives; *Semantic web description languages;* **•Theory of computation → Data provenance; •Applied computing → Document management and text processing;**

KEYWORDS

Data provenance, semantic web, digital libraries, metadata management, research workflows, text processing

JCDL '17, Toronto, Ontario, Canada
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

1 BACKGROUND

Workbench environments like the HathiTrust Research Center's (HTRC) Data Capsule service allow a researcher to analyze rights-restricted collections through support for a *non-consumptive research* paradigm. In this paradigm, a researcher engages in computational analysis by bringing their algorithms to the data in a manner that respects licensing arrangements for in-copyright works and exporting derivative data which does not reproduce these works. However, non-consumptive research can be difficult to understand in practice. We address this limitation here, through a conceptual framework to bridge the Capsule environment with their collection (*i.e.*, workset) through simple provenance-tracking workflows. This metadata contextualizes results and ensures that exports generated from Capsule-based research are non-consumptive.

The HathiTrust Digital Library (HTDL) of over 15 million digitized volumes from research libraries worldwide is a rich historical resource for research and scholarly investigation. The HathiTrust Research Center (HTRC) was formed in 2011 to help researchers formulate research questions and develop state of the art tools for computational analysis over this large heterogenous collection to support non-consumptive research. Non-consumptive use includes computational analysis of one or more volumes in the HTDL, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand the expressive content presented within that volume [2].

Non-consumptive use is supported in HTRC primarily through its Data Capsule service. Developed through a three year grant from the Alfred P. Sloan Foundation, the Data Capsule service provisions researchers with a Capsule, a virtual machine that runs in the trusted HTRC environment. The service guarantees non-consumptive use of the HathiTrust through a combination of policy and security checks [17, 22]. The central control is a toggle between allowing external Internet access and allowing access to the Capsule's secure volume. Derivative data are transferred out of the Capsule through a release spool that queues the files for HTRC review.

This paper extends the Data Capsule model with a conceptual framework to bridge the interaction with a user's collection (*i.e.*, workset), and demonstrates support for scholarly workflows with data provenance tools for non-consumptive research.

First, it describes the the conceptual underpinnings of the paper: Data Capsules, workflows, worksets, and data provenance. It

then discusses the implementation of these components in support of seamless analysis and publishing from a Capsule of the Data Capsule service. We conclude with a discussion of open questions regarding publication of analytical results from the Data Capsule service.

2 CONCEPTUAL COMPONENTS

Our experience has shown that a bare Capsule is too high a barrier for researchers who wish to work with HTDL materials. Even once a researcher uploads all their favorite text mining and visualization tools into their Capsule, users still exhibit a cognitive gap in how HTDL products and outputs relate to their analysis tools. We introduce workflows and worksets to address that gap, and show how data provenance tracking can help researchers reproduce their analysis, while ensuring production of non-consumptive exports.

2.1 Data Capsule

The Data Capsule service provisions virtual machines (VMs) to researchers within the HTRC secure environment. The VM and software environment together form a Capsule. Each researcher has exclusive use of the Capsule for a period of weeks or months during which they can configure their own environment for performing research on HTDL texts. More information on the Data Capsule service appears in [22] and [17].

2.2 Workflow

To explain workflows, we provide an case study of a non-consumptive workflow common to the digital humanities: topic modeling [1]. Scholars first create a collection of focal documents, then train a topic model on that collection to extract common themes. Topic models have been successfully applied in a number of digital humanities research activities involving both HTDL materials [5, 12, 20] and external materials [7, 19, 21].

The Capsule-based topic modeling workflow, previously demoed at JCDL 2015 [13], and summarized visually Figure 1, is made up of 6 steps:

(1) **Curate** workset from available HTDL materials,
(2) **Download** volumes to Capsule in secure mode,

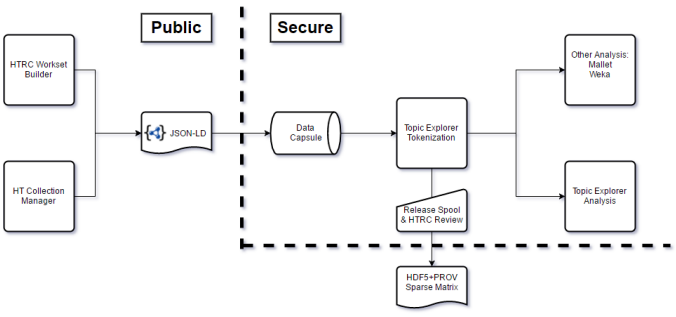

Figure 1: Topic modeling workflow. Public data is computed on researcher's desktop where secure data is computed within researcher's Capsule.

(3) **Tokenize** documents and apply stoplists to remove articles and prepositions. Remove common and rare words from corpus to better control dimensionality of LDA topic model,
(4) **Train** multiple topic models at variety of granularities, from 20-500 topics,
(5) **Visualize** results using InPhO Topic Explorer [11], and
(6) **Export** non-consumptive results, along with the tokenized and stoplisted corpus as structured data.

This workflow guarantees that the exported products are legitimate non-consumptive exports. The original corpus cannot be reconstructed from the exported corpus file, as stoplists have been applied destructively and the order of remaining words have been scrambled. This maintains reuse in bag of words models, but forbids direct reading of the in-copyright or rights-restricted text. The same transformations are applied to the public domain texts.

2.3 HTRC Worksets

Scholarly usage of digital libraries often involves locating texts to curate a collection, or workset, for further study. To have utility, these worksets must be a citable research product and afford flexibility to the preferred unit of analysis from the volume level up to the serial level or down to the page or even sentence level.

The HTRC Workset model resolves these issues of *addressability* and *relational expressivity*. In its simplest form the workset is the input dataset to an analysis task. In its fullest sense, it represents a researcher's intent from formation of a research question to the publication of research results.

In development of the workset, we carried out a comparison of four existing ontologies for bibliographic description. This revealed that no single surveyed ontology corresponded entirely to HathiTrust use cases [14]. In response, the HTRC Workset Ontology [6] contains key relations such as *htrc:intendedForUse*, *htrc:hasResearchMotivation*, *edm:isGatheredInto*, *dcterms:creator*, and *htrc:hasCriterion*.

2.4 Data Provenance

Data provenance is the lineage of a digital object. It captures factors that influenced an item or artifact's creation and transformation including the actors, agents (algorithms) and datasets that were part of the transformation of a digital object from one state to another. Data provenance is critical to the asserting the trustworthiness of a digital data product, especially when time and/or distance separate the digital object from its creator [10, 16].

Data provenance captured during use of a Capsule can aid a researcher in numerous ways. A researcher can interact with a Capsule over a period of weeks to months. Over that time, their workset may be continuously refined as content is added or removed. Provenance tracks these changes. As data analysis is carried out, new data products will be created, some of which will be exported as non-consumptive exports while others will be discarded as not of use. It is the role of provenance to track and connect this activity taking place over the duration of extended engagement of a researcher with their Capsule.

Over time, these analytical actions accumulate and non-consumptive exports are created. The publishable end product of research can then be conceived as the set $R = \{NC{-}exports, Workset, Provenance\}$

where $NC - export$ is one or more exports deemed to be non-consumptive, $Workset$ is the workset created and refined by the user, and *Provenance* is the record of activity that links $NC-export$, intermediate results, and the $Workset$.

3 MODEL OF INTEGRATION

The framework of workflow, workset, and data provenance within a Capsule provide a coherent way for researcher interaction both with their defined workset, and with the HTDL collection over which their analysis will be carried out.

A key component of the framework is a RESTful API through which a workset is imported into and exported from a Capsule. The workset is converted to a JSON-LD representation [18]. This API adheres to the Linked Data Platform (LDP) practices for publishing linked data [8]. LDP has been implemented in several digital libraries, including the OCLC's WorldCat [4], Digital Public Library of America[1], and Library of Congress[2]. Workflows are annotated with the PROV ontology [9], which defines a common vocabulary for describing data provenance. The components of this framework and their interaction is summarized in Figure 1.

Using the topic modeling workflow example from earlier, we created functionality to automatically create a provenance graph for the topic analysis using the PROV standard [3]. This provenance capture allows for proper attribution to the workset curator, to the researcher, and possibly to a third party running analysis algorithms on behalf of a research project.

Through use of the vocabulary of the PROV ontology [9], we describe a workset as a *prov:Collection*, declaring that each *htrc:Workset* is a *prov:Collection*. Each workflow, like the topic modeling workflow above, corresponds to a *prov:Plan*. The actual commands run by the analysis are a *prov:Activity*, while the artifacts created by an analysis are a *prov:Entity*. Each organization or person that contributes to an analysis is a *prov:Agent*.

Figure 2 shows an example provenance graph for the modeling stages of the workflow, following the plan *te:topicexplorer*. Three actions are carried out: *r1act*, *r2act*, and *r3act*, each corresponding to the tokenization, stoplisting, and training phases of the model, all of which are executed by a *user*. The underlying workset is abstracted as the *corpus* node. Files created by the tokenization and stoplisting process are specified by the *prov:wasGeneratedBy* and *prov:wasDerivedFrom* relations.

The training step (*r3act*) shows the highest level of sophistication in the graph and highlights the potential utility of provenance. The training produces three topic models, the abstract entity of which are labeled as *model10*, *model20*, and *model40*. The actual instantiated models are then tied to the training step by the *prov:wasGeneratedBy* relation, showing how this specific instance of a model was trained. Through the *r3act prov:Activity*, we can see that these models were trained specifically on the *r2* revision of the workset, which has been tokenized by *r1act* and stoplisted by *r2act*. Going further through the provenance graph associated with *corpus* would reveal the partner institutions and scanning practices that went into creating the workset that the user analyzed.

[1] https://dp.la/info/developers/map/
[2] http://ld.loc.gov/

All of this enables a quick way to validate and replicate results, tracking the software versions used to create the analysis and the command line arguments that were used. The PROV graph additionally aids in connecting non-consumptive exports to analysis activities and worksets, thus providing the fundamental principles and components needed for publishing of research done using the Data Capsule system.

4 CONCLUSION

Large collections of text will likely contain a mix of open-access, public-domain, rights-restricted, and in-copyright works. This should not mean that the whole collection must become a walled garden, off limits to researchers. The Data Capsule service, extended with a conceptual framework for user Worksets, is an answer to accessible but constrained access to large-scale restricted content. The selection and identification of corpora that are large subsets of a collection, something we call the "million volume workset", and its support within the Data Capsule environment is an ongoing challenge.

The next step for the development of the HTRC Data Capsule service is to stitch together each component to facilitate publication of a result of Capsule research. This step includes giving persistent identifiers to non-consumptive exports, and creating a published record that links together the non-consumptive exports, the provenance, and the workset.

Work is also needed to expand the provenance graph for worksets to include metadata about the scanning partner and the physical artifact. We hope to improve the precision of a workset description by using the FaBiO ontology [15] to distinguish between the use of a volume as a placeholder for an intellectual work (*i.e.*, we are pragmatically using a specific copy of Shakespeare's *The Tempest*, and any copy would fulfill the analysis) or as a particular manifestation (*i.e.*, this copy is the first folio edition, and it is important we use it specifically). This requires additional research in user experience to educate users about the subtle differences in usage patterns so they can properly encode their worksets.

ACKNOWLEDGMENTS

This work is funded in part through grants from the Andrew W. Mellon Foundation and the Alfred P. Sloan Foundation.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 4-5 (2003), 993–1022. DOI: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993
[2] Eleanor Dickson, Brandon Butler, Aaron Elkiss, Bobby Glushko, Robert McDonald, Sandra McIntyre, Leanne Nay, and Naz Pantaloni. 2017. *Non-Consumptive Use Research Policy.* Technical Report. https://www.hathitrust.org/htrc_ncup
[3] Yolanda Gil and Simon Miles. 2013. {*PROV*} *Model Primer.* {W3C} note. W3C.
[4] Carol Jean Godby, Shenghui Wang, and Jeffery K. Mixter. 2014. *Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description.* Morgan & Claypool.
[5] Andrew Goldstone and Ted Underwood. 2014. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History* 45, 3 (2014), 359–384. https://muse.jhu.edu/journals/new
[6] Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data* 2, e1 (2016). DOI: http://dx.doi.org/10.5334/johd.3
[7] M.L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press.

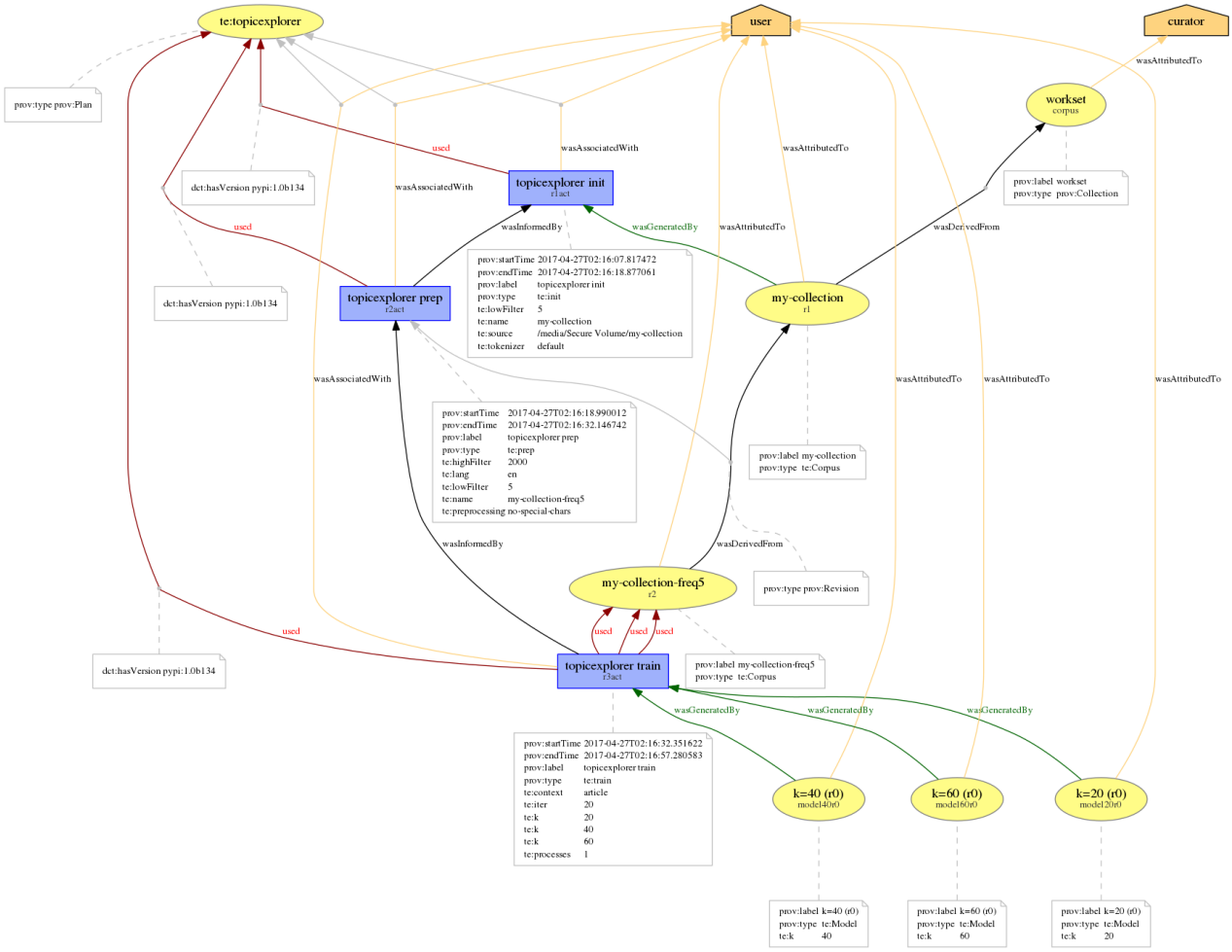

Figure 2: Sample provenance graph of topic modeling workflow. Actions displayed top-to-bottom in blue squares; research products as yellow ovals; users as orange pentagons. Graph follows visual conventions of the PROV Primer.

[8] Ashok Malhotra, Steve Speicher, and John Arwe. 2015. *Linked Data Platform 1.0.* {W3C} recommendation. W3C.

[9] Deborah McGuinness, Satya Sahoo, and Timothy Lebo. 2013. *PROV-O: The PROV Ontology.* {W3C} recommendation. W3C.

[10] Luc Moreau and Paul Groth. 2013. *Provenance: An Introduction to Prov.* Morgan & Claypool Publishers. http://www.provbook.org/

[11] Jaimie Murdock and Colin Allen. 2015. Visualization Techniques for Topic Model Checking. In *Proceedings of the 29th AAAI Conference (AAAI-15).* AAAI Press, Austin, TX. http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10007

[12] Jaimie Murdock, Colin Allen, and Simon DeDeo. 2017. Exploration and exploitation of Victorian science in Darwin's reading notebooks. *Cognition* 159 (feb 2017), 117–126. DOI: http://dx.doi.org/10.1016/j.cognition.2016.11.012

[13] Jaimie Murdock, Jiaan Zeng, and Robert H McDonald. 2015. Topic Exploration with the HTRC Data Capsule for Non-Consumptive Research. In *JCDL '15 Proceedings of the 15th ACM/IEEE-CS joint conference on Digital libraries.* ACM Press, Knoxville, Tennessee, USA. DOI: http://dx.doi.org/10.1145/2756406.2756929

[14] Terhi Nurmikko-Fuller, Kevin R. Page, Pip Willcox, Jacob Jett, Chris Maden, Timothy Cole, Colleen Fallaw, Megan Senseney, and J. Stephen Downie. 2015. Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15).* ACM, New York, NY, USA, 169–172. DOI: http://dx.doi.org/10.1145/2756406.2756944

[15] Silvio Peroni and David Shotton. 2012. FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web* 17 (2012), 33 – 43. DOI: http://dx.doi.org/10.1016/j.websem.2012.08.001

[16] Beth Plale, Inna Kouper, Allison Goodwell, and Isuru Suriarachchi. 2016. Trust Threads: Minimal Provenance for Data Publishing and Reuse. In *Big Data is Not a Monolith: Policies, Practices, and Problems*, Cassidy R. Sugimoto, Hamid Ekbia, and Michael Mattioli (Eds.). MIT University Press, Cambridge, MA.

[17] Beth Plale, Atul Prakash, and Robert McDonald. 2015. *The Data Capsule for Non-Consumptive Research: Final Report.* Technical Report. Indiana University. http://hdl.handle.net/2022/19277

[18] Manu Sporny, Markus Lanthaler, and Gregg Kellogg. 2014. *{JSON}-LD 1.0.* {W3C} recommendation. W3C.

[19] Timothy R. Tangherlini and Peter Leonard. 2013. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics* 41, 6 (2013), 725 – 749. DOI: http://dx.doi.org/10.1016/j.poetic.2013.08.002 Topic Models and the Cultural Sciences.

[20] Ted Underwood. 2015. The literary uses of high-dimensional space. *Big Data & Society* 2, 2 (2015), 2053951715602494. DOI: http://dx.doi.org/10.1177/2053951715602494

[21] Daniel Walker, Eric Ringger, and Kevin Seppi. 2013. Evaluating supervised topic models in the presence of OCR errors. *Proceedings of SPIE Document Recognition and Retrieval XX* 8658 (2013), 865812–865812–12. DOI: http://dx.doi.org/10.1117/12.2008345

[22] Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud Computing Data Capsules for Non-Consumptive Use of Texts. In *ScienceCloud '14: Proceedings of the 5th ACM Workshop on Scientific Cloud Computing.* Vancouver, BC, Canada, 9–16. DOI: http://dx.doi.org/10.1145/2608029.2608031