Research Center as Distant Publisher: Developing Non-Consumptive Compliant Open Data Worksets to Support New Modes of Inquiry

Robert H. McDonald-Indiana University, Bloomington, IN USA

July 10, 2016

Abstract

The HathiTrust Research Center (HTRC), founded in 2010, is managed by Indiana University Bloomington and the University of Illinois at Urbana-Champaign under an agreement with the HathiTrust Board of Governors and the University of Michigan. The HTRC mission supports new knowledge creation through novel computational uses of the Hathitrust Digital Library (HTDL). Through the introduction of the concept of *distant publishing*, this short paper will discuss ideas for data and software publication that support the HTRC non-consumptive research methodologies and offer scholars new methods for research inquiry.

1 Introduction

In the original Google Books Settlement Agreement in 2008 [1], funds were to be set aside to create a research center that would enable researchers worldwide to accomplish data-mining and analysis on texts in the public domain and under copyright in a manner that was secure and compliant with appropriate U.S. copyright law. This did not happen, because the court rejected the agreement in 2011. Despite this, in 2011, the HTDL announced that Indiana University Bloomington and the University of Illinois at Urbana-Champaign would run the HTRC under a cooperative funding agreement with the HathiTrust Board of Governors and the University of Michigan. Since 2014, HTRC has made available as an active production service tools to analyze a set of out-of-copyright content equaling around 4.4 million volumes. In 2016, the HTRC plans to enable analysis of the entirety of the 14 million volume corpus currently held by the HTDL, the largest digital academic library in North America.

2 HTRC and Non-Consumptive Research

The HTRC has developed a process to define and work within the concept of *non-consumptive* computational access to support the fair-use of the HTDL

corpus as defined within the Google Books Settlement Agreement that was a part of the *Authors Guild et al. v. Google Inc* case.

Currently the HTRC defines the process for *non-consumptive* use of the HTDL corpus as:

> Research in which computational analysis is performed on one or more books, but not research in which a researcher reads or displays.

Operationally, from the perspective of the HTRC research cyberinfrastructure, the HTRC defines *non-consumptive* research as:

> That which requires that no action or set of actions on the part of users, either acting alone or in cooperation with other users over the duration of one or multiple sessions can result in sufficient information gathered from a collection of copyrighted works to reassemble pages from the collection.

This concept has been further refined in the course of the development of the HTRC Data Capsule [12] for secure data analysis and the development of the HTRC Workset Ontology [5].

3 HTRC as Publisher

During the course of work with scholars using the HTRC tools and services to create derivative non-consumptive data sets, the Center has often taken on a set of the roles traditionally played by publishers. These data sets are reviewed by members of the HTRC staff for compliance with non-consumptive use standards prior to release to the authors.

As part of this work, the HTRC has offered as a service the capability to publish these non-consumptive, compliant data sets using a DOI scheme [2]. This service enables the creation of new derivatives [3] of published non-consumptive, compliant data sets.

A second benefit of opening access to these data sets is the ability to replicate current experiments that have been developed using the HTDL corpus and the HTRC tool set. From this standpoint the HTRC functions as a *distant publisher* of non-consumptive compliant data sets in support of new models of research inquiry.

4 Distant Publishing as Concept

Prior to defining the concept of *distant publishing*, it is first instructive to understand *distant reading* within the context of digital humanities. *Distant reading* was first codified in 2000 by noted humanist and scholar Franco Moretti:

> Distant reading: where distance . . . is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, less is more. [7]

Moretti later expanded the concept in his 2013 monograph of the same name [8].

Much like Moretti's definition that focuses on enabling a broader view of the text, the *distant publisher* enables a broader view of data sets through bringing to bear the current corpus of computational tools for large-scale textual data mining and analysis. HTRC as a distant publisher is removed by at least one degree from the creator, and remains distinct from any standardized concept of publisher. Yet, data sets are published under the rubric of the HTRC, and these publications are freed from the constraints of copyright in this context due to their non-consumptive nature. Thus we define distant publishing as

> Publication of a non-consumptive data set outside of any standardized publishing construct, removed by x degree from the original creator,
> openly available to the community of scholars for replication and available for re-use in support of the advancement of knowledge.

This definition is one that the HTRC aims to further refine in the coming years. We welcome broader thoughts on this concept from those working to preserve open research data and the software that makes that data accessible for use in scientific experimental replication and re-use for the long-term benefit of the scholarly community.

5 Distant Publishing Use Cases

Currently the HTRC is developing models that define the notion of *distant publishing*. These models are illustrated in several use cases, outlined below.

- **Extracted Features Worksets** - HTRC expects this concept to be further refined as we move toward the second round of HTRC Advanced Collaborative Support grants which will be funded in summer 2016. Our most progressive case for distant publishing at this point is leveraged through the publication and release of our main extracted features workset. The current workset is a prototype based on the 4.8 million volume public domain collection from the HTDL. Through 2016-17 this workset will be redefined to include more of the HTDL collection. From this initial workset publication we have seen further refinements of the workset by scholars such as Ted Underwood [10], Colin Allen [9], and Matthew Wilkens [11].

- **HT+Bookworm** - The HathiTrust+Bookworm (HT+BW) project [6] presents textual content through interactive visualization. Whereas HT+BW has previously been used in standalone contexts with pre-determined metadata, currently HT+BW is enabling scholars to analyze custom personal collections from within the larger corpus and the use of HT+BW as a supplement to other uses of the HTRC. This concept could eventually become a new possibility for derived workset publication in its own right.
- **HTRC Workset Ontology** - Currently in development, the HTRC Workset Ontology is part of a collections data model by the Workset Creation for Scholarly Analysis project [4], a HTRC research initiative funded by the Andrew W. Mellon Foundation. The resulting HTRC Workset data model is designed to aid humanities scholars by helping them to describe selected portions of the HTDL corpus that serve as the objects of their research. The resulting worksets are persistent, citable, and can be assessed by other scholars for reuse in additional research processes.

6 Conclusion

Today's digital scholars are embracing new opportunities to explore their disciplines through the type of enhanced computational analysis that the HTRC provides. As the Center works to define emerging possibilities within the context of non-consumptive research, distant publishing will enable us to engage with the community of open data and open software publishers to ensure that our collections are accessible, open and available for the next generation of distant readers and their plans for new forms of scholarship.

7 Acknowledgment

The author would like to thank the Executive Leadership Team of the HTRC, J. Stephen Downie, Beth A. Plale, Beth Naymachchivaya, and John M. Unsworth, and all of the staff of the HathiTrust Research Center and the HathiTrust Digital Library for their contributions to the tools and services that make the concepts in this paper possible.

8 License

References

[1] Paul N Courant. The Stakes in the Google Book Search Settlement. *The Economists' Voice*, 6(9), jan 2009.

[2] Boris Capitanu; Ted Underwood; Peter Organisciak; Sayan Bhattacharyya; Loretta Auvil; Colleen Fallaw; J. Stephen Downie;. Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes, 2015.

[3] Ted Underwood; Boris Capitanu; Peter Organisciak; Sayan Bhattacharyya; Loretta Auvil; Colleen Fallaw; J. Stephen Downie;. Word Frequencies in English-Language Literature 1700-1922 (0.2), 2015.

[4] HTRC. Workset Creation for Scholarly Analysis - A HATHITRUST RESEARCH CENTER PROJECT FUNDED BY THE ANDREW W. MELLON FOUNDATION, 2016.

[5] Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data*, 2, mar 2016.

[6] P. Organisciak L. Unnikrishnan B. Schmidt M. Shamim R.H. McDonald J. Downie E. Aiden L. Auvil, S. Bhattacharyya. Adding Flexibility to Large-Scale Text Visualization with HathiTrust+Bookworm. In *Proceedings of Digital Humanities 2016*, pages 854–56. Alliance of Digital Humanities Organizations, 2016.

[7] Franco Moretti. Conjectures on World Literature. *New Left Review*, 1:57–58, jan-feb 2000.

[8] Franco Moretti. *Distant Reading*. Verso, 2013.

[9] Jaimie Murdock, Jiaan Zeng, and Colin Allen. Towards Cultural-Scale Models of Full Text. *Arxiv.org*, 2016.

[10] Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu. Mapping mutable genres in structurally complex volumes. In *2013 IEEE International Conference on Big Data*. Institute of Electrical & Electronics Engineers (IEEE), oct 2013.

[11] Matthew Wilkens. Literary Geography at Corpus Scale. In *Proceedings of Digital Humanities 2013*. Alliance of Digital Humanities Organizations, 2013.

[12] Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing - ScienceCloud '14*. Association for Computing Machinery (ACM), 2014.