Creating a Policy Framework for Analytic Access to In-Copyright Works for Non-Consumptive Research

Eleanor Dickson
dicksone@illinois.edu
University of Illinois, United States of America

Daniel G. Tracy
dtracy@illinois.edu
University of Illinois, United States of America

Sandra McIntrye
mcintsan@hathitrust.org
HathiTrust Operation, United States of America

Bobby Glushko
rglushko@uwo.ca
University of Western Ontario, Canada

Robert H. McDonald
rhmcdona@indiana.edu
Indiana University, United States of America

Brandon Butler
bcb4y@eservices.virginia.edu
University of Virginia, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois, United States of America

Introduction

We report on the work of a recent HathiTrust Research Center (HTRC) task force charged to draft an actionable, definitional Non-Consumptive Use Research Policy. As the research division of HathiTrust, the HTRC facilitates computational text analysis of materials in the HathiTrust Digital Library (HTDL) by adhering to a non-consumptive research paradigm. As the HTRC has integrated the text of the full HTDL corpus into its datastore, it has become increasingly important to clarify and codify the Center's policy for non-consumptive research. The task force, which consisted of copyright and scholarly communications librarians and representatives from HathiTrust operations and the HTRC, recommended a policy that clarifies acceptable researcher behavior and allowable exports from the HTRC Data Capsule (Plale, et al., 2015). This poster describes the task force's work to establish a Non-Consumptive Use Research Policy for the HTRC that aims to achieve the same goals as copyright itself: to promote progress in the discovery and spread of knowledge, without harming the commercial interests of authors, publishers, and other stakeholders.

Background

While the concept of non-consumptive research has seeded the mission of the HTRC, the Non-Consumptive Use Research Policy task force sought to translate conceptual definitions of the term into practicable policy (Bhattacharyya, et al., 2015). When first used in 2010, the term was defined as "research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book" (Amended Settlement Agreement: Authors Guild, Inc., et al., v Google Inc., 2009). Since then, legal scholar Matthew Sag and literary scholar Matthew Jockers have offered their own definitions and assessments, tending to favor instead the term *non-expressive use* (Sag, 2012; Jockers, 2013). Several recent court decisions pointed the task force toward the current legal understanding of non-consumptive research specifically and current interpretations of fair use broadly (Authors Guild v HathiTrust, 2014; Cambridge University Press v Becker, 2016; Fox News Network v TV Eyes, 2014). Additionally, the task force looked to existing access models for restricted data (ICPSR) as well as professional guidelines for non-consumptive research (Association of Research Libraries, 2012; Cox, 2015).

Policy Highlights

The group first created a framework drawing on fair use that, when paired with the HTRC technical infrastructure, would clarify non-consumptive access to the HTDL. This framework accounted for several considerations and safeguards:

- Mechanical data mining differs from researcher-driven computational text analysis, which requires interplay between scholar and text.
- Current case law suggests that it needs to be sufficiently difficult, but not strictly impossible, to reconstruct the expressive work (Authors Guild v Google Books, 2015).
- Users must agree that they will not treat HTRC tools as a reading application, and the tools should periodically remind them of this limitation.
- The HTRC must continue to block through technological measures and human review the export of protected textual data from the secure system.

The task force then drafted the HTRC Non-Consumptive Use Research Policy (HathiTrust, 2017). It defines non-consumptive research as "Research in which computational analysis is performed on one or more volumes or textual objects in the HTDL, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted work to understand the expressive content presented within that work." Of key importance is the notion of substantial portion, which, according to the policy, is a portion of the work sufficient in quality or quantity to provide a substitute for access to the expressive content of the original text. The policy outlines acceptable in-capsule uses of corpus text that are limited to those which would facilitate scholarly text analysis, including checking results to refine algorithms. In addition to enumerating non-consumptive research practices—for example text extraction, textual analysis, and automated translation—the policy provides sample results that further model approved uses. These results, which may be exported from the HTRC Data Capsule, include non-binary, human-readable statistical summaries, derived results, keywords-in-context, and concordances that are not sufficient to reconstruct a substantial portion of the text.

Conclusions

The task force tailored the policy to address current infrastructure within the HTRC, both technical and human, as opposed to accounting for prospective updates to interface and design. As such, the policy is an iterable, living document that must be revisited as HTRC systems are further developed. Such technical developments, such as the HTRC's exploration of machine-aided results verification to augment the current human-review system, will improve the scalability of the HTRC Data Capsule and may require updates to the policy. As more researchers interact with the HTRC Data Capsule, their use cases may prompt additional refinement of the policy, especially in the exemplar results it provides. The process followed in developing the policy, as well as the guidelines themselves, may be useful in other text mining research environments. They encourage an interpretation of non-consumptive research that values scholarship and intellectual progress, while still balancing the restrictions imposed by copyright law.

Acknowledgements

Task force members included: Aaron Elkiss, Brandon Butler, Bobby Glushko, Daniel G. Tracy, Eleanor Dickson, Robert McDonald, and Sandra McIntyre.

Bibliography

Amended Settlement Agreement: Authors Guild, Inc., et al., v Google Inc. (2009).

Association of Research Libraries (2012). *Code of Best Practices in Fair Use for Academic and Research Libraries.* Available from: http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf (accessed 1 November 2016).

Authors Guild v Google Books (2015).

Authors Guild v HathiTrust (2014).

Bhattacharyya, S., Organisciak, P., and Downie, J. S. (2015). "A fragmentising interface to a large corpus of digitized text." *Interdisciplinary Science Reviews,* 40(1): 61-77.

Cambridge University Press v Becker (2016).

Cox, K. L. (2015). "ARL Issue Brief: Text and Data Mining and Fair Use in the United States," Available from: http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf (accessed 1 November 2016).

ICPSR (2017). "Data Enclaves," *University of Michigan*, Available from: http://www.icpsr.umich.edu/icpsrweb/content/icpsr/access/restricted/enclave.html (accessed 1 November 2016)/

Fox News Network v TV Eyes (2014).

Jockers, M. (2013). *Macroanalysis: digital methods and literary history.* Champaign: University of Illinois Press.

HathiTrust (2017). "Nonconsumptive Use Research Policy." Available from: https://www.hathitrust.org/htrc_ncup (accessed 17 March 2017).

Plale, B., Prakash, A., and McDonald, R. (2015). "The Data Capsule for Non-Consumptive Research: Final Report."

Sag, M. (2012). "Orphan Works as Grist for the Data Mill." *The Berkeley Technology Law Journal* 27: 1503-50.