All for One and One for All: Reconciling Research and Production Values at the HathiTrust Through User-Scripting

David Bainbridge
University of Waikato
Hamilton, New Zealand
Email: davidb@waikato.ac.nz

J. Stephen Downie
University of Illinois
Urbana-Champaign, USA
Email: jdownie@illinois.edu

***Abstract*—This article details a practical technique that safely reconciles the production stability and integrity of the HathiTrust Digital Library (HTDL) with the riskier and potentially disruptive experimental functionalities created by the HathiTrust Research Center. Web systems produced by HTRC are necessarily more speculative and, understandably, operate on equipment outside of the HTDL production environment. The key to our approach that brings these two parts closer together is to exploit user-scripting: a web browser add-in technique that allows users to introduce bespoke Javascript code that alters the behavior of specific website(s). We demonstrate how it can be used to provide a mashup of three web sites: HTDL and two web-based offerings operated independently by HTRC. The end result is that the user interacts with the HTDL as usual, and at strategic locations in the interface additionally functionality drawn from the research systems—which takes account of the user's current context—is seamlessly blended in.**

I. Introduction

The HathiTrust is a consortium of institutions and libraries committed to "the long-term curation and availability of the cultural record."[1] Founded in 2008, its membership—which is open to institutions worldwide—has grown to over 100 organizations. A key technological element to the HathiTrust is the development and on-going support of a digital library (hereafter HTDL) providing access to the increasing body of material that members (and commercial partners such as Google) have digitized and applied Optical Character Recognition (OCR) to, drawn from their libraries and archives. At the time of writing HTDL included 14,888,528 volumes (books, serials etc.) comprising of in excess of 5.2 billion pages, requiring some 667 Terabytes of disk storage. It is managed by the University of Michigan.

The default for a newly digitized work is to assume the work is in copyright, which corresponds to restricted access in HTDL. When an item has been confirmed as being in the public domain its metadata record is updated, corresponding to full-access in the digital library. The HathiTrust also provides a suite of Application Program Interfaces (APIs) and exported files for download, again aligned with copyright information. Approximately 60% of HTDL content is under copyright restrictions.

The HathiTrust *Research Center* (HTRC) is the research arm of the HathiTrust. It is dedicated to "facilitating scholarship by enabling analytic access to the corpus, developing research tools, fostering research projects and communities, and providing additional resources such as enhanced metadata and indices that will assist scholars to more easily exploit the HathiTrust materials" [3]. Founded in 2011, HTRC is co-located at the University of Illinois and Indiana University.

In previous work we have shown the viability of user-scripting [2] in digital library systems to provide a more flexible approach to dealing with author name variants which can confound a user's task of locating relevant work [1]. In this new work we show how it can be used to modify a user's interactive experience of the main HTDL to include functionality drawn from related, but externally managed, experimental systems. We demonstrate utilizing two research systems at the HTRC: the Bookworm project[2] and the Extracted Feature Dataset,[3] which we respectively elaborate on over the next two sections.

II. Bookworm Splice

In the HTRC Bookworm project unigrams of all the words in all the volumes in HTDL are stored in such a way that when a user enters a word (or series of words) a time-line based visualization of the frequency of that (or those) words in all the volumes is produced.

It is a separate website that a user needs to know about to make use of. Consider now Figure 1 where our user has the HTDL + HTRC mashup user-script installed in their browser.[4] When the user visits the home page to HTDL, instead of there being just one button for search (with a magnifying glass on it) located to the right of the main search box, there are now two buttons. The second button shows the graphic of a cartoon worm on it, and when this is pressed the terms

[1] https://www.hathitrust.org/community

[2] https://analytics.hathitrust.org/bookworm

[3] https://analytics.hathitrust.org/datasets

[4] Available through http://data.analytics.hathitrust.org/features/ and is compatible with all major web browsers.

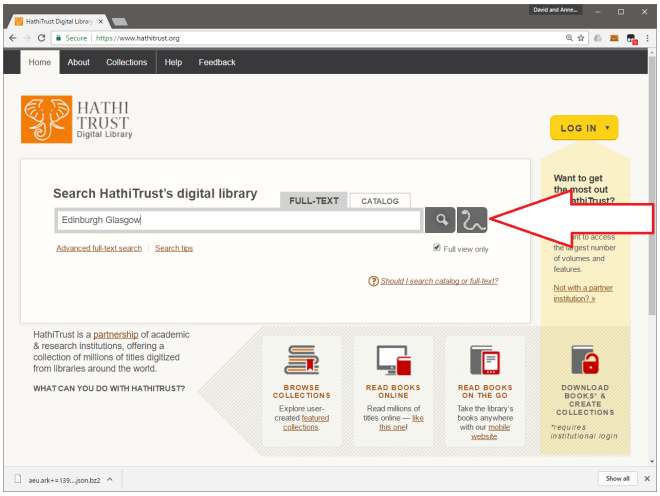

Fig. 1. Bookworm search button splice.

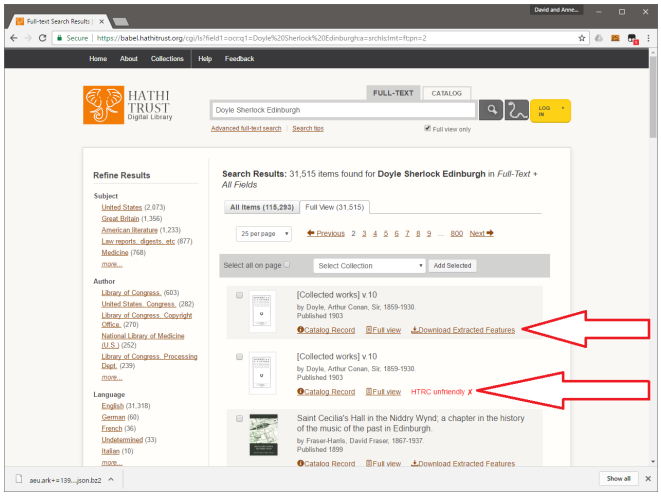

Fig. 2. Augmented search results list, providing a directly link to download extracted features when available.

entered into the search box are used to initiate directly a bookworm search with those terms, instead of performing the normal search function. The inclusion of this button is sympathetic with the overall interface—we have augmented the figure with a prominent arrow to help identify its location. Entering the terms *Edinburgh* and *Glasgow* and then pressing the bookworm button, for example, results in the browser displaying (via the HTRC bookworm website) a plot over time of the frequency of occurrences of these two words in the books stored in HTDL.

III. Extracted Feature Integration

The HTRC Extracted Feature Dataset takes the idea of unigram information utilized in the Bookworm project further, providing a JSON file per volume that quantifies the frequency of each unique word on each page. This is augmented with additional information, such as part-of-speech tags per word and language identification per page, as derived by processing the OCR'd text with OpenNLP.[5]

One of the aims of the Extracted Feature Dataset is to allow scholars to download the JSON files for the volumes they are interested in, and run analytical software—whatever is relevant to them—over it. Figure 2 shows an additional benefit the user-scripting approach can provide to the status quo. While searching and browsing HTDL, an additional link is added per displayed item, allowing the user to directly download the Extracted Feature JSON file for that volume. Again, prominent arrows have been used to highlight where such changes occur. Although it does not happen very often, it can be the case that the volume the user finds was recently added to the digital library and has not yet been added to the Extracted Feature Dataset. This disconnect is not readily apparent in the standard interface. With the user-script active, an AJAX call is made to make this determination and items found to be missing in the Extracted Feature Dataset are flagged as such. The example in the figure given was carefully selected so an example of this could also be seen, marked in red.

HTDL also allows users to specify and publish collections of volumes, whose metadata can then be downloaded. Although not shown in the figure, the user-script adds an additional button for converting a HathiTrust collection into an HTRC Extracted Feature workset. The same volume check for existence in the HTRC dataset is made, and volumes that do not yet exist in the dataset are marked as missing, a task that would ordinarily be non-trivial to perform by the scholar.

IV. Conclusions

In conclusion, the developed user-script provides an immediate gain for scholars seeking access to the resources provided by the HathiTrust, allowing easier access to distributed resources; moreover it delivers context-based functionality that is highly complimentary with what is currently provided by the HathiTrust and its Research Center. Indeed the approach taken can be generalized to allow the HathiTrust to assess the merits of various research-led innovations without major disruptions to their production services, from which more informed plans to incorporate (or not) the work into the production environment can be drawn up.

References

[1] D. Bainbridge, M. B. Twidale, and D. M. Nichols, "Interactive context-aware user-driven metadata correction in digital libraries," *Int. J. Digit. Libr.*, vol. 13, no. 1, pp. 17–32, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1007/s00799-012-0100-5

[2] A. Boodman and J. Dunck, "Greasemonkey," *Firefox browser extension, http://greasemonkey. mozdev. org*, 2007.

[3] J. S. Downie, "The HathiTrust research center: Providing analytic access to the HathiTrust digital library's 4.7 billion pages," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '15. New York, NY, USA: ACM, 2015, pp. 5–5.

[5] https://opennlp.apache.org