Author Manuscript

TORCHLITE: New, Open Analytical Tools and Infrastructure for a Mega-Scale Digital Library

Lamba, Manika HathiTrust Research Center (HTRC), University of Illinois, USA | manika@illinois.edu
Walsh, John HTRC, Indiana University, USA | jawalsh@indiana.edu
Dubnicek, Ryan HTRC, University of Illinois, USA | rdubnic2@illinois.edu
Christie, Jennifer HTRC, Indiana University, USA | jechri@iu.edu
Downie, J. Stephen HTRC, University of Illinois, USA | jdownie@illinois.edu
Swatscheno, Janet HTRC & HathiTrust, University of Michigan, USA | jswatsch@hathitrust.org
Kudeki, Deren HTRC, University of Illinois, USA | dkudeki@illinois.edu
Layne-Worthey, Glen HTRC, University of Illinois, USA | gworthey@illinois.edu

ABSTRACT
This paper introduces TORCHLITE, an innovative HathiTrust Research Center (HTRC) open analytical and computational framework designed to offer efficient, open, and approachable access to the HTRC Extracted Features (EF) dataset via a well-documented web-based API. This poster will summarize project goals and progress, and discuss community engagement, which has played a pivotal role in this project. During a hackathon event held in spring 2024, TORCHLITE fostered collaboration among digital humanities and information science scholars to develop widgets and notebooks utilizing the EF API. Through the hackathon, participants explored the API's capabilities, leading to the creation of over a dozen analytical widgets and interactive programming notebooks (e.g., Jupyter).

KEYWORDS
HathiTrust, Digital Libraries, Community Engagement, API, Data Visualization

INTRODUCTION
TORCHLITE is a new open, beginner-friendly analytical and computational framework from HathiTrust Research Center (HTRC) providing efficient programmatic access to the Extracted Features (EF) dataset via a well-documented, web-based API (Walsh, 2023). It utilizes a non-relational (NoSQL) database tailored for hosting the EF dataset (Jett et al., 2020), and empowers users to extract highly specific subsets of EF data, down to individual volumes or pages. The API can be used to supply data to new, user-friendly, and user-created analytical tools and visualizations (which we informally call "widgets") and includes development of a public data dashboard to host widgets for the analysis of volumes and collections (worksets) from the HathiTrust Digital Library (HTDL). With the open API and public data dashboard, users can share widgets with anyone, as well as host widgets in third-party environments such as library catalogs, blogs, etc., or run and share them as standalone interactive programming notebooks (e.g., Colab, Jupyter, or Observable notebooks) or web applications. The TORCHLITE public dashboard was iteratively developed based on input gathered from the user community in the form of early-stage focus groups and ongoing user tests and currently supports a number of widgets for analysis of "*worksets*" (computable collections of texts selected from the 18.5 million-volume HathiTrust Digital Library). Currently available widgets include:

- *Summary:* Summary statistics about document length and vocabulary density;
- *Simple Word Cloud:* Word-frequency visualization and statistics;
- *Publication Date Timeline:* Timeline visualization of publication dates of volumes in the workset;
- *Mapping Contributor Data:* Map-based visualization that plots the birthplaces of authors represented in the workset.

Figure 1 shows examples of web-based widgets that were built using open data visualization libraries (e.g., *Chart.Js, D3.js*). The widgets, dashboard, and derived notebooks suit a broad audience, including users without programming skills. All code developed as part of TORCHLITE is made available under an open-source license, and we have developed and promoted these tools and API through a robust community outreach program, culminating most prominently in a community hackathon, focused on widget and notebook building.

TORCHLITE API AND DASHBOARD
TORCHLITE API is a RESTful API that creates and retrieves EF data specific to user-, community-, and expert-defined worksets. The TORCHLITE Dashboard was built using Mantis, a React.js template with aesthetics derived from Ant Design (2024) and components used from MUI (2024). It is built using the following technologies and APIs:

- Next.js-based front-end interface (what the user sees in their browser)

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pra2.1160 This article is protected by copyright. All rights reserved.

- Next.js backend API (the server-side piece of the Next.js application)
- TORCHLITE API (implements the TORCHLITE business logic for dashboards, filters, and data management)
- Extracted Features API (used for accessing Extracted Features data for specific volumes and worksets)
- Registry API (used for retrieving public and user-private worksets)

TORCHLITE HACKATHON: Community Engagement Event

We promoted the EF data and related APIs, stress-tested infrastructure and generated prototype widgets and widget ideas in a hackathon event held from 21-23 May, 2024 at the University of Illinois at Urbana-Champaign. 50 participants, including HathiTrust Research Center staff, worked in different teams to produce nearly a dozen widgets and notebooks. The hackathon brought together digital humanities, information science, and computer science scholars, library developers, and independent researchers to develop tools and visualizations using the TORCHLITE APIs. Supported by workshop tutorials on the API and HTRC data, attendees were also invited to lead open-source projects in novel ways. Projects were shared with peers and HTRC affiliates for feedback and will be considered for eventual implementation in the widget dashboard. This hackathon increased public computational engagement with one of the largest digital library collections in existence, and provided invaluable testing, feedback, and brainstorming on TORCHLITE design and functionality. Additionally, the hackathon allowed HTRC staff to conduct multiple user tests that will drive evolution of the TORCHLITE interface through project end and beyond.

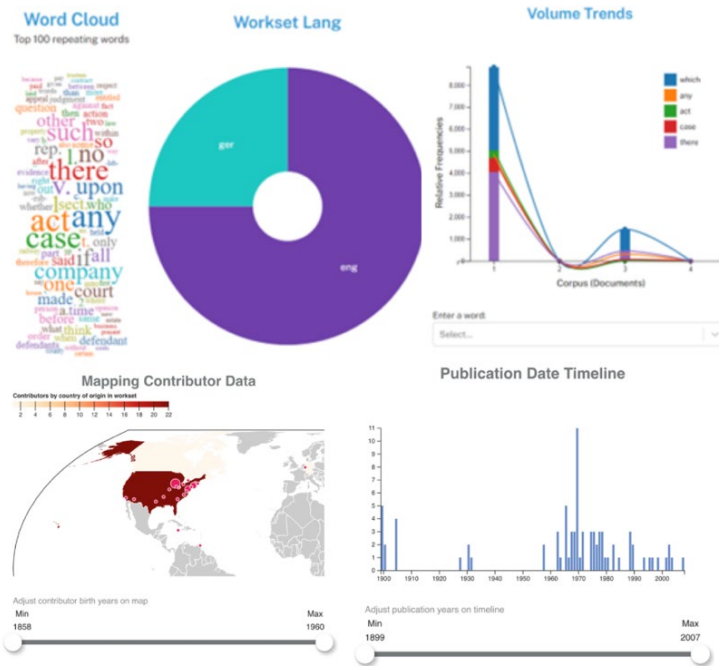


Figure 1. Example Widgets Generated using EF API

CONCLUSION

The extracted features (EF) 2.0 dataset contains nearly 3 trillion tokens representing over 6 billion pages of digitized text from dozens of academic libraries around the world, making it arguably the largest open dataset of its kind readily available to information science and other scholars in the world (Jett et al., 2020). The EF 2.0 dataset includes a JSON representation of the complete text of all these pages, presented as copyright-compliant (i.e., non-consumptive) data enhanced with a variety of statistical and linguistic measures obtained through a natural language processing pipeline, and includes library catalog data as linked data-compliant BIBFRAME. However, the breadth and richness of this dataset have presented one significant downside: it is large (over 4.2 TB compressed) and unwieldy.

Our poster introduces TORCHLITE, a more user-friendly and accessible framework for accessing, analyzing, and presenting this rich dataset in ways that will allow researchers and developers with a variety of technical skills to interact with meaningful subsets of this open data. TORCHLITE's open framework allows for a potentially endless variety of analytical methods, visualizations, and applications to be created with code libraries of the user's choice, using a variety of computational frameworks (embedded and standalone widgets, code notebooks, etc.) that can consume data via API. We anticipate that this open, flexible mode of access will greatly increase the usability and use of this community-created dataset.

Author Manuscript

ACKNOWLEDGMENTS

The TORCHLITE project is generously funded by the U.S. National Endowment for the Humanities (Grant no. HAA-284850-22). Views, findings, conclusions, and recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities.

GENERATIVE AI USE

We confirm that we did not use generative AI tools/services to author this submission.

AUTHOR ATTRIBUTION

Walsh, Downie, Layne-Worthey: conceptualization, methodology, formal analysis, writing – review and editing; Lamba, Dubnicek: writing – original draft, data curation, formal analysis, visualization; Christie, Kudeki, Swatscheno: project administration, formal analysis, methodology, writing – review and editing.

REFERENCES

Ant Design—The world's second most popular React UI framework. (n.d.). Retrieved May 28, 2024, from https://ant.design/

EF | Extracted Features API. (n.d.). Retrieved May 28, 2024, from https://htrc.stoplight.io/docs/ef-api/06db4dc572b49-ef

Jett, J., Capitanu, B., Kudeki, D., Cole, T., Hu, Y., Organisciak, P., Underwood, T., Dickson Koehl, E., Dubnicek, R., & Downie, J. S. (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)* [dataset]. HathiTrust Research Center. https://doi.org/10.13012/R2TE-C227

MUI: The React component library you always wanted. (n.d.). Retrieved May 28, 2024, from https://mui.com/

torchlite-hackathon. (n.d.). TORCHLITE Hackathon. Retrieved May 28, 2024, from https://htrc.github.io/torchlite-hackathon/

Walsh, J., Layne-Worthey, G., Jett, J., Capitanu, B., Organisciak, P., Dubnicek, R., & Downie, J.S. (2023). "The library is open!": Open data and an open API for the HathiTrust Digital Library. In CHR2023: Computational Humanities Research Conference.

Welcome to HathiTrust | HathiTrust Digital Library. (n.d.). Retrieved May 28, 2024, from https://www.hathitrust.org/about/

Author Manuscript