Journal of Open Humanities Data

By the People Crowdsourcing Datasets from the Library of Congress

DATA PAPER

VICTORIA VAN HYNING
LAUREN ALGEE
MASON JONES
CARLYN OSBORN
TREVOR OWENS
LAUREN SEROKA
ABBY SHELTON

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The *By the People* (*BTP*) datasets comprise text of selected collections of the Library of Congress (LOC) created by volunteers in the *By the People* crowdsourced transcription program, which invites public transcription of historical documents. All transcriptions are created and reviewed by volunteers in a consensus-based model in which two or more volunteers must agree on a transcription for it to be considered complete. Resulting transcriptions are added to the digital collections alongside the images to enable search and accessibility of the collections. Additionally, completed transcription "campaigns" are published as freely downloadable datasets of .CSV files containing all campaign transcriptions, as well as minimal metadata. The datasets can support a multitude of purposes including computational research in fields such as history, linguistics, economics, and political science.

CORRESPONDING AUTHOR:
Victoria Van Hyning
College of Information Studies, University of Maryland, College Park, USA
vvh@umd.edu

KEYWORDS:
crowdsourcing; text; transcription; cultural heritage; accessibility; Handwritten Text Recognition

TO CITE THIS ARTICLE:
Van Hyning, V., Algee, L., Jones, M., Osborn, C., Owens, T., Seroka, L., & Shelton, A. (2022). *By the People* Crowdsourcing Datasets from the Library of Congress. *Journal of Open Humanities Data*, 8: 5, pp. 1–10. DOI: *https://doi.org/10.5334/johd.67*

(1) OVERVIEW

REPOSITORY LOCATION

Library of Congress, Washington, DC, USA. All *By the People* campaign datasets are available at this url as they are published: *https://www.loc.gov/search/?fa=contributor:by+the+people+%28program%29*. The seven datasets available at time of publication are detailed below.

CONTEXT

The *By the People* (*BTP*) datasets comprise text of selected collections of the Library of Congress (LOC) created by volunteers in the crowdsourced transcription program, which invites public transcription of historical documents. *BTP* has two indivisible goals:

1. Engage virtual volunteers in meaningful opportunities to deeply explore and learn from the LOC's collections.
2. Enhance discoverability and usability of LOC collections through volunteer-created text (Shelton, 2021).

As of 1 November 2021, *By the People* contributors have completed transcriptions for 392,000 images, and transcribed an additional 102,000 images which await peer review. Monthly updates are available at *https://crowd.loc.gov/about/*. Completed transcriptions are published alongside the images in *loc.gov* to enable image-level word search and improve readability and accessibility of the collections. These image-level transcriptions are also individually downloadable as .TXT files. Additionally, text from completed transcription "campaigns" are published in bulk as downloadable .CSV datasets. To date, seven completed campaigns, including a total of 23,316 transcriptions, have been published as datasets for bulk download, and these are the primary subject of this paper. These seven campaigns can be found in ***Table 1***.

Table 1 Dataset Description.[1]

DATASET	SUMMARY OF CAMPAIGN	DATASET URL	LCCN	OBJECT NAME	NUM. OF TRANSCRIPTIONS
Wm. Oland Bourne Papers	Selection from the papers of reformer, poet, editor, and clergyman William Oland Bourne (1819–1901). Includes narratives submitted by disabled Union veterans in a Left-hand Penmanship contest sponsored by Bourne as well as Civil War reminiscences by soldiers and sailors in Central Park Hospital, New York, N.Y.	*https://www.loc.gov/item/2019667237*	*https://lccn.loc.gov/2019667237*	civil-war-soldiers-disabled-but-not-disheartened_2020-12-10.csv	5,159
Branch Rickey Papers	Selections from the papers of Branch Rickey, major league baseball manager and executive, consisting of scouting reports from the 1950s and 1960s. They are mostly concentrated in the years 1951–1956 and 1962–1963, while Rickey was associated, respectively, with the Pittsburgh Pirates and St. Louis Cardinals.	*https://www.loc.gov/item/2019667234/*	*https://lccn.loc.gov/2019667234*	branch-rickey-scouting-reports	1,926

(Contd.)

1 This table includes all datasets available at the time of publication.

DATASET	SUMMARY OF CAMPAIGN	DATASET URL	LCCN	OBJECT NAME	NUM. OF TRANSCRIPTIONS
Samuel J. Gibson Diary and Correspondence	The papers of Union soldier Samuel J. Gibson (1833–1878) consist of a letter and diary written by Gibson in 1864 while serving with Company B, 103rd Pennsylvania Infantry Regiment, and as a prisoner at Camp Sumter in Georgia, the Confederate prisoner of war camp commonly known as Andersonville Prison.	*https://www.loc.gov/item/2019667238/*	*https://lccn.loc.gov/2019667238*	hell-upon-earth-gibson_2020-12-10.csv	90
Carrie Chapman Catt Papers	The papers of suffragist, political strategist, and pacifist Carrie Lane Chapman Catt (1859–1947) span the years 1848–1950, with the bulk of the material dating from 1890 to 1920. The collection consists of approximately 9,500 items (11,851 images), most of which were digitized from 18 microfilm reels.	*https://www.loc.gov/item/2019667239/*	*https://lccn.loc.gov/2019667239*	carrie-chapman-catt-papers-2020-12-07.csv	5,760
Elizabeth Cady Stanton Papers	The papers of suffragist, reformer, and feminist theorist Elizabeth Cady Stanton (1815–1902) cover the years 1814 to 1946, with most of the material concentrated between 1840 and 1902.	*https://www.loc.gov/item/2020445592*	*https://lccn.loc.gov/2020445592*	elizabeth-cady-stanton-papers-2021-04-19.csv	3,456
Rosa Parks Papers	Selections from the papers of Rosa Parks (1913–2005), including personal and family correspondence, personal writings and reflections, and ephemera from her speaking engagements and honors.	*https://www.loc.gov/item/2020445590*	*https://lccn.loc.gov/2020445590*	rosa-parks-in-her-own-words-2021-04-19.csv	1,769
Susan B. Anthony Papers	The papers of reformer and suffragist Susan B. Anthony (1820–1906) span the period 1846–1934 with the bulk of the material dating from 1846 to 1906.	*https://www.loc.gov/item/2020445591*	*https://lccn.loc.gov/2020445591*	susan-b-anthony-papers-2021-04-19.csv	5,156

From October 2018 to October 2021 the *BTP* team launched 24 thematic campaigns of content for the public to transcribe. Three campaigns for staff contribution have also launched since March 2020. These comprise over half a million images from LOC collections across four curatorial Divisions including Manuscript, American Folklife Center, Law Library of Congress, and Rare Book and Special Collections. Materials reflect the breadth and depth of the Library's collections and include selections from the personal papers of American presidents; archives of women's suffrage, civil rights, and abolition activists; writings of military leaders and veterans; papers of ethnomusicologist Alan Lomax; baseball scouting reports of Branch Rickey; legal documents; literary drafts of Walt Whitman; and records of occult experimentation from Harry Houdini's collection. Many of the materials are handwritten, but many campaigns also include typed text, most of which is not amenable to extraction via optical character recognition (OCR).

By the People launched on 24 October 2018 as a pilot of LC Labs, the LOC digital innovation unit, and in January 2020 became a permanent program of the Library's Digital Content Management Section. A team of three *By the People* community managers run the program

and support and encourage volunteers through newsletters, a discussion forum,[2] Twitter,[3] email, and virtual and in-person events including challenges and transcribe-a-thons.[4] Extensive "How-to" instructions[5] for transcription, review and tagging are available in English and Spanish.

BTP and the underlying open source codebase Concordia[6] are developed iteratively, and integrate user-centered changes in response to formal and informal user research and feedback and the requirements of different collection materials (Ferriter et al, 2019). The transcription conventions have not changed substantially, but additional guidance has been added to assist transcription of specific types of text, such as cross-writing.

(2) METHOD

STEPS

All transcriptions are created and reviewed by volunteers using a consensus model, in which at least two volunteers must agree on a transcription for it to be marked as complete. All activity takes place online at *crowd.loc.gov*. ***Image 1*** depicts a page that has been transcribed and reviewed on the *BTP* interface. Anyone can transcribe without an account. Registered users can review other volunteers' transcriptions, add tags, and access their contribution history. Volunteers browse to select assets to transcribe or review. Participants can see all pages in sequential order and their current status — "Not started", "In Progress", "Needs Review", or "Completed."

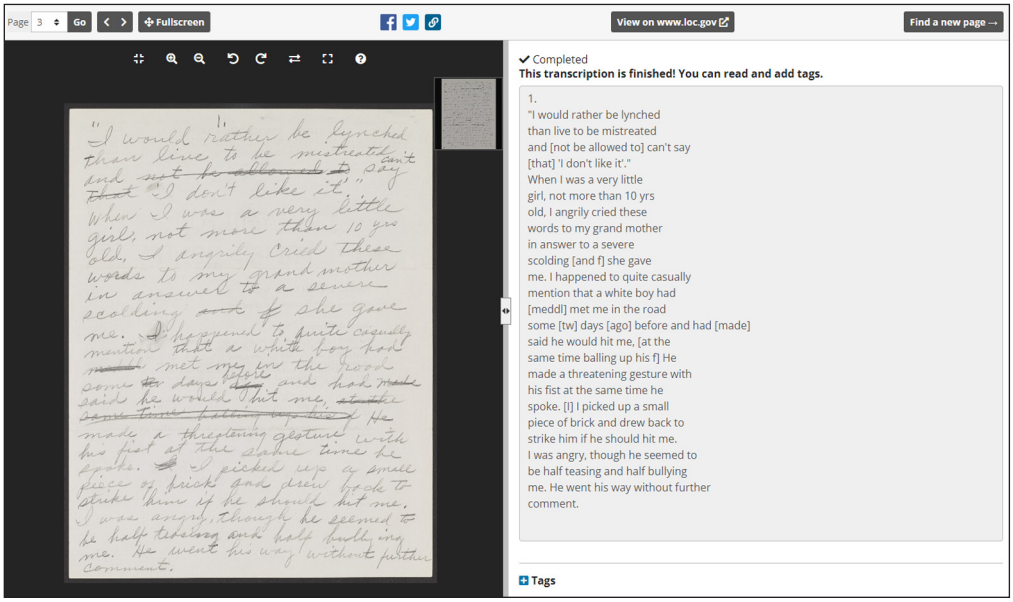

Image 1 Screenshot of a completed transcription from the Rosa Parks Papers in By the People.[7]

Volunteers are asked to preserve all original spelling, punctuation, and line breaks, except in cases where words break over lines or pages. These conventions create transcriptions that are word searchable, amenable to screen reader technology, and a good starting point for those wishing to use the data for handwritten text recognition training systems. To make the data most useful for Handwritten Text Recognition (HTR), a user would have to go through each transcription and edit words breaking over lines or pages to reflect the original layout of the page.

Images are imported into *BTP* via the *loc.gov* Application Programming Interface (API). A *loc.gov* item, consisting of one or more images, is also called an "item" in *By the People*, while the individual images are called "assets." Assets may show more than one page of documents, though in *BTP* outreach assets are often referred to colloquially as "pages."

2 *https://historyhub.history.gov/community/crowd-loc* Accessed November 22, 2021.

3 LOC Crowdsourcing, @Crowd_LOC. Accessed November 22, 2021. *https://twitter.com/Crowd_LOC*.

4 *https://crowd.loc.gov/resources/* Accessed November 22, 2021.

5 *https://crowd.loc.gov/help-center/welcome-guide/* Accessed November 22, 2021.

6 *https://github.com/LibraryOfCongress/concordia* Accessed November 22, 2021.

7 *https://crowd.loc.gov/campaigns/rosa-parks-in-her-own-words/writings-notes-and-statements/mss859430227/mss859430227-3/*, [Webpage]. Accessed November 22, 2021.

Campaigns consist of chronological or thematic buckets of items called "projects". The campaign content may come from a single collection or unite materials across collections, as in the case of Walt Whitman.[8] Projects can also be linked to Topics, connecting related content across campaigns. Current topics include the Civil War, presidential papers, and women's suffrage.

BTP is a pass-through application that does not automatically sync with *loc.gov*. Changes made to images or text in one will not automatically appear in the other. Once transcriptions go through quality control (see below) they are manually exported from *BTP*, ingested into long-term preservation storage,[9] published as part of the items on *loc.gov*, and packaged as datasets. An attribution is included at the end of each transcription's .TXT file: "Transcribed and reviewed by contributors participating in the *By the People* project at *crowd.loc.gov*" (Van Hyning, 2020).

QUALITY CONTROL

LOC subject specialists spot-check completed campaigns before publication of the transcriptions for quality control, as is exemplified in the completed campaigns included in ***Table 1***. Often their review begins when the campaign does, so that early interventions can be made in the instructions or campaign context if a repeated error is spotted. The number of assets checked varies campaign-to-campaign and is determined by the specialist. During their review before the data is exported, the specialists edit significant errors they encounter such as mistranscribed words that change the meaning of the text; or supply missing words, phrases or (in rare cases) whole pages. Very few instances of vandalism have occurred to date. At least five percent of all currently available datasets were reviewed by LOC staff; Gibson and Parks were reviewed in their entirety. Two additional quality control assessments have been undertaken and are summarized below.

The Branch Rickey papers contain 1,926 pages of baseball scouting reports; some are memos, others tabular. The materials are relatively human-legible, but OCR did not meet curatorial needs due to thin paper, often faint type, and other issues. Algee et al (2019) sampled 240 characters from the midpoint of all narrative Rickey transcriptions and found 98% accuracy. Typical errors included volunteers correcting spelling (against the *BTP* conventions), introducing typos, or not conforming to *BTP* formatting conventions.

In 2020, Manuscript Division subject specialist Michelle Krowl analyzed all 90 images of Samuel J. Gibson's diary transcribed in the "This Hell-upon-earth of a prison" campaign and logged all of her changes in a spreadsheet. The handwriting and spelling are representative of other nineteenth-century materials included in *BTP*. Krowl identified 703 character-level errors in total. Most were minor: expansions, such as changing "&" to "and"; typos or misspellings by volunteers; and correction of original spellings. Parts of some pages are so damaged or ambiguous that a definitive reading cannot be provided by the volunteers or specialists, and were not calculated in the error rate. One major source of error was an untranscribed page (281 characters) marked as "Complete." The overall character error rate for Gibson before Krowl made edits was 703/152,017 or .0046%.

Table 2 presents an excerpt of the Rosa Parks data and abridged transcriptions. We provide the Rosa Parks[10] dataset description here as a representative example of the seven currently available datasets, and template for future releases:

> This dataset includes: .ZIP file containing a .CSV file and a README file. - rosa-parks-in-her-own-words-2021-04-19.csv- a .CSV containing campaign, project, item, itemID, asset, and asset status metadata, as well as an image link, and the volunteer-generated transcription. This .CSV is the direct export of the "Rosa Parks: In Her Own Words" campaign.

8 *https://crowd.loc.gov/campaigns/walt-whitman/*, [Webpage]. Accessed November 22, 2021.

9 For more information on Library of Congress digital collection management practices see *https://www.loc.gov/programs/digital-collections-management/about-this-program/*, [Webpage]. Accessed November 22, 2021. *https://www.loc.gov/item/2020445590* [Webpage]. Accessed November 22, 2021.

10 *https://www.loc.gov/item/2020445590* [Webpage]. Accessed November 22, 2021.

Table 2 Rosa Parks .CSV File Excerpt. This table demonstrates the file structure and content.

CAMPAIGN	PROJECT	ITEM	ITEMID	ASSET	AssetStatus	DownloadURL	TRANSCRIPTION
Rosa Parks: In Her Own Words	Writings, Notes, and Statements	Rosa Parks Papers: Writings, Notes, and Statements, 1956–1998; Notebooks; 1961–1962 , 1985–1990, undated	mss859430232	mss859430232 -45	completed	*http://tile.loc.aov/imaae-services/iiif/service:mss:mss85943:0018:16:0045/full/oct:50/0/default.jpg*	Wednesday 1/25/89, 1PM Vivian's death Funeral service Sat. 10AM 1/28/89 Union Baptist Church Newfield Ave. Home Address, 53 Bonner Street Stamford
Rosa Parks: In Her Own Words	Writings, Notes, and Statements	Rosa Parks Papers: Writings, Notes, and Statements, 1956–1998; Notebooks; 1961–1962 ,	mss859430232	mss859430232 -44	completed	*http://tile.loc.aov/imaae-services/iiif/service:mss:mss85943:0018:16:0044/full/oct:50/0/default.jpg*	1/26/89 Star Makers Kelly & Co Agents Cheryl Kagan Mich Bank
Rosa Parks: In Her Own Words	Writings, Notes, and Statements	Rosa Parks Papers: Writings, Notes, and Statements, 1956–1998; Notebooks; 1961–1962 .	mss859430232	mss859430232 -43	completed	*http://tile.loc.aov/imaae-services/iiif/service:mss:mss85943:0018:16:0043/full/oct:100/0/default.jpg*	Dr. Anderson 24535 N. Carolina Southfield Ml 48075 met her at Farmer
Rosa Parks: In Her Own Words	Writings, Notes, and Statements	Rosa Parks Papers: Writings, Notes, and Statements, 1956–1998; Notebooks; 1961–1962 , 1985–1990, undated	mss859430232	mss859430232 -42	completed	*http://tile.loc.aov/imaae-services/iiif/service:mss:mss85943:0018:16:0042/full/oct:50/0/default.jpg*	New Bride + Groom Mr. & Mrs. Anderson Bowles Rev. Bynum's grand daughter Kelly 745–4795 have be there at a . a r

The README file provides more detailed information about each data field. These descriptions appear in the "Summary" field for each dataset. CSV files are all structured in the same manner, and named according to this formula: Campaign-name-YYYY-MM-DD.csv (e.g. carrie-chapman-catt-papers-2020-11-16.csv). The Branch Rickey datasets include an additional version of the data (V1) with two additional .CSVs in which the data are sorted by document format based on the categories established for the quality analysis described above. The workflow for transcription and dataset publication is outlined in ***Figure 1***.

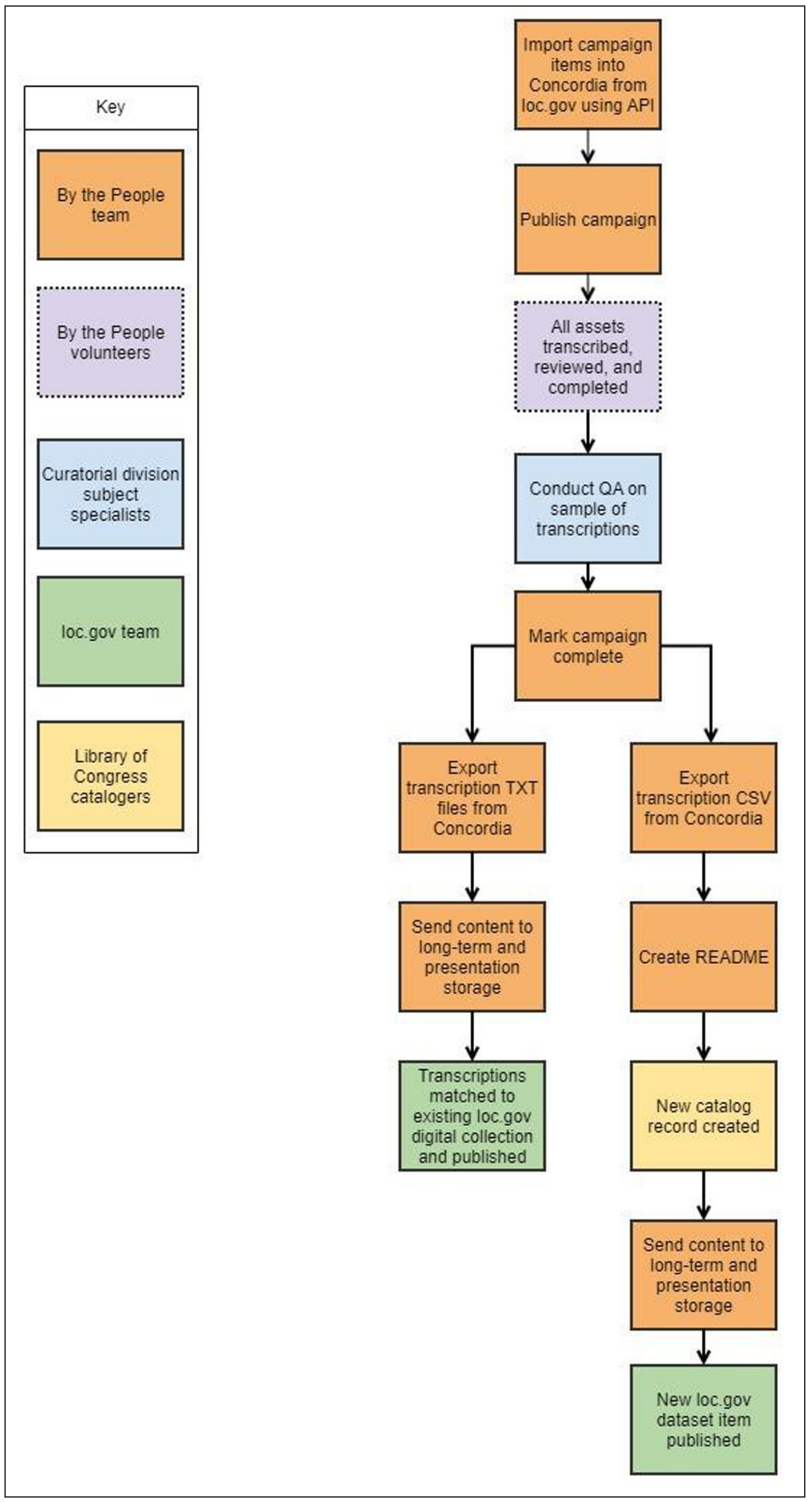

Figure 1 Transcription workflow diagram: The flow of transcription data from creation to publication and the different constituents who make this work possible.

FORMAT NAMES AND VERSIONS

README and .CSV.

CREATION DATES

2018-10-24 – 2021-04-19

DATASET CREATORS

The content of the datasets are the results of transcription and review by anonymous and registered *By the People* volunteers, with minor intervention by Library of Congress staff, including the *By the People* team.

LANGUAGE

English is the primary language for each of the current datasets, though they contain small amounts of other languages.

LICENSE

The Rights and Access information on the dataset page refers generally to Library of Congress-published datasets overall. The README provides the following information specific to the *By the People* datasets: "All contributions to the By the People application are released into the public domain as they are created. Anyone is free to use and re-use the datasets."

REPOSITORY NAME

Library of Congress

PUBLICATION DATE

- Rickey V1, 2019-03-22, Rickey V2 2020-06-16
- Bourne, 2020-12-10
- Gibson, 2020-12-10
- Catt, 2020-11-16
- Parks, 2021-04-19
- Anthony, 2021-04-19
- Stanton, 2021-04-19

(4) REUSE POTENTIAL

In addition to powering search and accessibility on *loc.gov*, *BTP* transcriptions have many potential research uses. Algee et al. (2019) modeled possibilities using Voyant tools on the Rickey dataset, including word clouds and semantic analysis. The authors found that while Rickey's most frequently used word was "good", the usage was most often critical of a player's abilities, as in "I think [Bob Wakefield] is a good man to get rid of" (30). Computational linguistics, including semantic, sentiment, and word frequency analysis, are well-supported by the data, as are traditional close-reading practices in the humanities (Seroka and Shelton 2021). Speeches, diaries, and other personal writings in the suffragist and civil rights papers are ripe for deeper study, while the extensive body of letters in most collections, particularly among the interrelated suffragists, would lend themselves to new network analyses.

These datasets offer significant opportunities to study the accuracy and quality of crowdsourced transcriptions, and could be used in combination with the *BTP* transcription conventions and scaffolding, *BTP* discussion forum analysis, user-surveys, and other qualitative and quantitative approaches to probe the efficacy of platform design and community engagement work in helping people learn about history, primary source use, paleogeography and more. Finally, these data have clear potential as training sets for improving machine-learning, HTR of manuscripts, and OCR of various materials.

ACKNOWLEDGEMENTS

Support for the collections, engagement, and technology of *BTP* is the result of extensive collaboration between the *BTP* team, LC Labs, IT Design and Development, the Digital Collections Management and Services Division, collection curators, and many others working across the Library of Congress.

FUNDING INFORMATION

Staffing for *BTP* and the development of the Concordia platform is supported in part by the National Digital Library Trust Fund.

COMPETING INTERESTS

Victoria Van Hyning serves on the editorial board of *JOHD*. Mason Jones serves as a copy editor for *JOHD*.

AUTHOR CONTRIBUTIONS

- Victoria Van Hyning (corresponding author): Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Writing – original draft.
- Lauren Algee: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Supervision, Visualization, Writing – original draft.
- Mason Jones: Writing – review & editing
- Carlyn Osborn: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Writing – review & editing
- Trevor Owens: Project administration, Resources, Supervision
- Lauren Seroka: Conceptualization, Software, Data Curation, Investigation, Methodology, Project Administration, Validation, Writing – review & editing
- Abigail Shelton: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Writing – review & editing

AUTHOR AFFILIATIONS

Victoria Van Hyning *orcid.org/0000-0001-6775-2870*
College of Information Studies, University of Maryland, College Park, USA

Lauren Algee
Digital Content Management Section, Library of Congress, Washington, DC, USA

Mason Jones *orcid.org/0000-0002-8777-2315*
College of Information Studies, University of Maryland, College Park, USA

Carlyn Osborn
Digital Content Management Section, Library of Congress, Washington, DC, USA

Trevor Owens *orcid.org/0000-0001-8857-388X*
Digital Content Management Section, Library of Congress, Washington, DC, USA

Lauren Seroka
Digital Content Management Section, Library of Congress, Washington, DC, USA

Abby Shelton
Digital Content Management Section, Library of Congress, Washington, DC, USA

REFERENCES

Algee, L., Ferriter, M., & **Van Hyning, V.** (2019). "'And the Crowd Goes Wild!': Crowdsourcing Baseball History at the Library of Congress." *Archival Outlook*, 4–5, 30, *https://mydigitalpublication.com/publication/?i=623810&article_id=3494347&view=articleBrowser*

Ferriter, M., Zwaard, K., Kamlley, E., Storey, R., Adams, C., Algee, L., Van Hyning, V., Bresner, J., Potter, A., Jakeway, E., & **Brunton, D.** (2019). "With One Heart": Agile approaches for developing Concordia and crowdsourcing at the Library of Congress. *The Code4Lib Journal*, 46. *https://journal.code4lib.org/articles/14901*

Seroka, L., & **Shelton, A.** (2021, June 10). "Diving into Branch Rickey: Using a dataset of crowdsourced transcriptions as a tool for open research". *The Signal* [Webpage]. Accessed November 22, 2021. *https://blogs.loc.gov/thesignal/2021/06/diving-into-branch-rickey/*

Shelton, A. (2021, October 21). "Using Crowdsourced Lincoln Transcriptions: An Interview with Jon White". *The Signal* [Webpage]. Accessed November 22, 2021. *https://blogs.loc.gov/thesignal/2021/10/jon-white/*

Van Hyning, V. (2020, July 9). "Finding By the People Transcriptions in the Library's Digital Collections". *The Signal* [Webpage]. Accessed November 22, 2021. *https://blogs.loc.gov/thesignal/2020/07/finding-by-the-people-transcriptions-in-the-librarys-digital-collections/*

DOI: 10.5334/johd.67

TO CITE THIS ARTICLE:
Van Hyning, V., Algee, L., Jones, M., Osborn, C., Owens, T., Seroka, L., & Shelton, A. (2022). By the People Crowdsourcing Datasets from the Library of Congress. *Journal of Open Humanities Data*, 8: 5, pp. 1–10. DOI: *https://doi.org/10.5334/johd.67*

Published: 04 February 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.