Crowdsourced Data: Accuracy, Accessibility, Authority (CDAAA)

Assistant Professor Victoria Van Hyning (PI) of the University of Maryland, College of Information Studies is requesting $446,525 to pursue a 3-year Early Career Research Development program to investigate 3 crucial aspects of cultural heritage crowdsourced transcription projects: a) data quality, b) the challenges of incorporating transcription data into the Content Management Systems (CMSs) that enable collections discovery and, c) the accessibility of these data for people who are blind, dyslexic or live with other disabilities that necessitate the use of screen readers--software that reads website content aloud. Online crowdsourcing projects and platforms have proliferated within GLAMs and universities since 2010, engaging millions of internet-connected volunteers around the world. Many GLAMs explicitly invite volunteers to help make their collections more discoverable and accessible, but later encounter unexpected challenges to data integration, which prevents or limits accessibility, and potentially wastes volunteers' time. Transcriptions are the most common type of crowdsourcing data solicited by GLAMs: when integrated into CMSs they broaden access to otherwise non-machine readable images of handwritten documents, and fill significant gaps in traditional archival description. GLAMs urgently need more information about the potential barriers to data integration and the real-world experiences of users, as well as roadmaps to success at crowdsourcing. This work is strongly aligned with LB21 Goal 2: Objective 2.3.

Crowdsourced Data: Accuracy, Accessibility, Authority (*CDAAA*) will pursue 4 interlocking research questions centered on the challenges of incorporating volunteers' transcription data into authoritative GLAM CMSs): **RQ1:** How much does transcription data quality vary across crowdsourcing projects and platforms? Data quality will be examined in terms of a) fidelity to project transcription conventions, b) character accuracy rate (comparable with OCR data quality measures) c) whether crowdsourcing platforms introduce errors and complexity **RQ2:** What are practitioners' beliefs about crowdsourced data quality, and do these align with the findings of RQ1? **RQ3:** When GLAMs successfully integrate crowdsourced data with CMSs, what methods and resources do they use? When they struggle, what are the barriers to success? What is needed to overcome barriers? **RQ4:** Is transcription data integrated with CMSs accessible for screen-reader users, and if not what is required to make the data legible?

Project Justification: Crowdsourced transcription platforms use different methods to gather data. Some methods may produce more accurate data than others or require more technical skill to clean and integrate into CMSs. Zooniverse has deployed several methods requiring multiple independent users to transcribe each page and algorithms to compare and aggregate the results, but aggregation introduces significant errors, requiring considerable data cleaning effort. Previous IMLS-funded Zooniverse work supported our development of a more accurate alternative method in which multiple volunteers transcribe, but everyone can see each other's transcriptions (Blickhan et al. 2019). What is not known is how this method compares to those of From the Page (FtP), NARA, Smithsonian, and Library of Congress projects, which all enable users to transcribe and edit each other's work without using algorithms to combine transcriptions. Which methods produce the most accurate datasets? Are some platforms better suited to certain document types? Are some data outputs more accessible to screen-reader users? In short, what platform should a GLAM use? GLAM practitioners and scholars are asking these questions, but there is little work that helps them make informed decisions.

Project Work Plan: Van Hyning will lead a PhD student affiliated with the Recovering and Reusing Archival Data (RRAD) Lab at UMD, which she co-founded in 2021, in a blended qualitative and quantitative approach. **Y1**: We will gather 15 crowdsourced transcription datasets from partner institutions, inc. FtP users: Folger, Driskell Center, UMD Special Collections (SCUA), others indentified by the Advisory Board (AB); Zooniverse users: Folger, Getty, UPenn (*Scribes of the Cairo Geniza*), Boston Public Library (*Anti-Slavery Manuscripts*), *Old Weather* partners (NARA, Royal Museums, Greenwich), and publicly available datasets from NARA, Smithsonian, and LOC, which use bespoke crowdsourcing platforms and CMSs. The data is diverse in terms of document layout, languages, time period, and subject matter, ranging from 10thC Hebrew manuscript fragments, to 16thC recipes, to 19th-20thC ship logbooks, to 20thC-21stC civil rights leaders' papers, including those of Rosa Parks (LOC). We will then deploy a new hybrid method to compare data accuracy from different projects and platforms. Drawing on Causer and Terras (2014), we will combine character error rate and convention violations (human error) with machine-introduced error (algorithm failures) adapted and

expanded from Blickhan et al. (2019). We will also conduct a Qualtrics survey of 60 GLAM practitioners (using SPSS for analysis), including our partners and others identified by the PI's, partners' and AB's networks. The survey will focus on: if and how respondents integrate crowdsourced data with CMSs and their beliefs about the quality of these data. **Y2**: We will do deeper semi-structured interviews of 20 survey participants about their crowdsourced data integration practices, challenges, and beliefs about data quality (10 project partners, 10 subjects identified through Y1 survey). Interviews will be recorded, transcribed through Rev.com, and coded using grounded theory methods and NVivo software (aligning coding via inter-rater reliability testing) by the PI and PhD. Working with a postdoctoral fellow from UMD iSchool's Trace Center, we will recruit 12 screen-reader users to test accessibility of transcriptions in partners' CMSs. Drawing on User-Centered Design methods (where 12 users is industry standard), we will conduct video-recorded contextual interviews to observe users navigating CMSs to a) attempt to find a document with transcribed text and b) read it with their screen-reader. We will code usability barriers and successes per Web Content Accessibility Standards 2.0. **Y3**: We will complete the interview and UX accessibility interview coding and disseminate findings through publications and conferences.

Project Director and Partnerships: With 8 years of industry and research experience in crowdsourcing, and an extensive publication and outreach record, Van Hyning is uniquely well-placed to lead this project. As a postdoctoral fellow and Humanities PI of Zooniverse (2014-18) she pioneered new text transcription and data aggregation methods for *Shakespeare's World* (Folger Shakespeare Library), *AnnoTate* (Tate) and other GLAMs. For the IMLS-funded "Transforming Libraries and Archives through Crowdsourcing" grant (2015-19) she contributed to new transcription methods for *Scribes of the Cairo Geniza* and *Anti-Slavery Manuscripts*, and co-conducted the first study to assess the accuracy of Zooniverse transcription methods (Blickhan et al. 2019). As a Senior Innovation Specialist at the Library of Congress, she co-created *By the People* (2018-20) and led the effort to return transcriptions to the CMS. She also uses FtP crowdsourcing software in her teaching and research, i.e. *David C. Driskell Papers Project* (UMD). **Project partners** are the PI's existing collaborators. They use one or more crowdsourcing platforms, and a range of off-the-shelf and bespoke CMSs: Folger (Zooniverse, FtP, and bespoke transcription platform Dromio; bespoke CMS); *Old Weather* (7 Zooniverse platform iterations; GLAM partners inc. NARA, Royal Museums, Greenwich and others); the Driskell Center (FtP; Past Perfect CMS); SCUA (FtP, bespoke crowdsourcing platform, Fedora, ArchiveSpace); other Zooniverse users (Boston Public Library, UPenn, Getty, Huntington). **Confirmed Advisory Board members include:** Zuhair Mahmood, Blind accessibility tester and expert (formerly of LOC); Samantha Blickhan (Zooniverse, Adler Planetarium), Mark Matienzo (Assistant Director for Digital Strategy and Access, Stanford Libraries), and Ben Brumfield (co-creator of FtP).

Diversity plan: A core goal of this work is to evaluate the accessibility of crowdsourced transcriptions in CMSs for screen-reader-users with a range of disabilities i.e. blindness, severe epilepsy, dyslexia. Project partners include public and private GLAMs, and projects that increase access to English-language 19thC-21stC BIPOC authors, artists, and thinkers, women writers from the 16thC-21stC, and other languages including Hebrew.

Project Results: We will disseminate our findings via conferences such as ASIS&T, Citizen Science Association, ALA, SAA, and open access research publications i.e. *Citizen Science Theory and Practice*, *Journal of Open Humanities Data* and *JASIST* in Y3. Cross-project and platform data analysis will be published as datasets on GitHub under a CC BY-NC-SA 4.0 license. The data and publications will enable practitioners to compare data quality, project design, and data collection methods from different projects for the first time. These data will also provide partners and fellow practitioners with concrete suggestions for accessibility improvements.

Budget: The PI requests $446,525 for this work: $178,584 ($95,159 in stipends, $26,454 health coverage, $56,971 tuition) for a 12-month doctoral student GA for 3 years; 1 course release Y1-Y3 for Van Hyning ($34,616), and 1 month of summer salary per year ($38,424), inc. fringe benefits ($13,240); Y2 .15 FTE of Trace Center postdoctoral researcher to recruit screen-reader users, co-conduct usability testing and disseminate research findings ($17,500); $4,225 for participant incentives; $4,000 for office supplies and a laptop for the GA; $10,000 for open access publishing fees, conference registration and travel for team members Y1-Y3 to disseminate findings. $3,750 for 5 AB members ($250/yr/3 yrs); and negotiated 54.5% indirect cost ($137,415).