Crowdsourcing, Zooniverse, and Archives

Svea Williams

School of Library and Information Science, The University of Southern Mississippi

LIS 661: Archival Capstone

Dr. Fay

December 3rd, 2025

Crowdsourcing refers to the concept of utilizing the efforts of volunteers to accomplish a task that would otherwise take an individual or a smaller group much longer to complete. This practice is particularly useful for archives, given that many institutions have large amounts of information that are inaccessible to the public for a variety of reasons. Over the past two decades, virtual crowdsourcing has become a well-respected method for unlocking a wealth of information that would only be accessible with a large investment of manpower, which many institutions would not be able to afford without the help of volunteers. The goal of this literature review is to chronologically examine a selection of scholarly works written about the benefits of crowdsourcing projects for archives, with a particular focus on projects that utilize the Zooniverse platform.

General Scholarship

Published in 2012, Estellés-Arolas and González-Ladrón-de-Guevara sought to establish a comprehensive definition of the term crowdsourcing as it applies to information science. The authors compared 166 papers from various scholarly sources in order to gather information, and then broke the definition into three parts (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). The authors first defined the crowd as a group of individuals working to complete tasks of varying levels of difficulty. The crowd benefits from this interaction in a variety of ways, most often through the receipt of monetary payments, social recognition, or the development of new skills; in short, the authors concluded that participants have one of Maslow's needs fulfilled (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). The host organization benefits from this collaboration by allowing the crowd to solve its problem through its collective intelligence. In the case of a crowdfunding project, the host organization also benefits from the crowd's monetary assets. The authors concluded that by the very nature of a crowdsourcing or

crowdfunding project, the entire interaction takes place online. This article provides a helpful foundation for examining additional literature related to crowdsourcing projects, especially in understanding the basic motivations of participants.

Ellis's 2014 publication examines how crowdsourcing and collaboration can specifically benefit British libraries and archives. The author explains that the term "has been used in diverse ways to describe the process of using large groups of people to meet a need, either through user-generated (UG) content, UG research, transcription, editing" (Ellis, 2014). Ellis continues by emphasizing the importance of utilizing the crowd's collective intelligence to make progress on the backlog of projects held by cultural heritage institutions. While some information professionals may feel threatened by allowing the untrained public to perform various functions of their jobs, it also gives the community a chance to engage with and shape collection processes, the terminology used, and the user experience. The authors emphasize the importance of ensuring that volunteers feel valued and can see the impact of their contributions in order to encourage continued engagement. It is also noted that any user-generated content that does not utilize a controlled vocabulary must be evaluated periodically to ensure the language remains accessible in the future. Given the United Kingdom's history with crowdsourced projects such as the Oxford English Dictionary, it is easy to see how similar modern efforts may become equally popular.

In their 2017 article, Alam and Campbell analyzed the motivations of volunteers participating in the Australian Newspaper Digitization Program (ANDP) organized by the National Library of Australia. The goal of the study was to analyze the intrinsic and extrinsic motivations of participants to determine how they benefited from their volunteer efforts. The ANDP enables volunteers to correct optical character recognition (OCR) transcriptions of

newspaper text line by line or by article at their discretion. The platform also allows volunteers to tag articles, add comments, and participate in forum discussions. Alam and Campbell interviewed twelve participants who engaged with the project in various ways. Among the interviewees were retired individuals seeking to pass the time, authors conducting research, and people focused on information about their families in particular. Overall, they found that participants' motivations were altruistic and focused on preserving and improving access to Australia's history. Based on their study, the authors concluded that citizen scientists are typically driven by intrinsic factors such as their own personal interests or feelings of social responsibility, but may also be encouraged by common goals and a sense of community (Alam & Campbell, 2017).

Iranowska's 2019 article examines how a crowdsourcing project's user interface affects its success and popularity. The author considers other important factors in project participation, including volunteer motivation, effective interface design, and a positive user experience. Iranowska then compares Zooniverse projects as a whole to two projects hosted on independent websites, Edvard Munch's Writings and Transcribe Bentham, both of which are based on the MediaWiki platform. She begins with the homepage of each project and notes that both MediaWiki projects have very text-heavy homepages that lack any call to action, and the Munch website is only available in Norwegian or German. While Zooniverse emphasizes the importance of every contribution, the MediaWiki projects focus on the accuracy of each transcription, given that each resource is only examined by one volunteer. Munch's Letters and Transcribe Bentham both offer project leaderboards and point systems, while Zooniverse designers emphasize the importance of creating an ideal interface. Finally, Iranowska explains that hosting a project on MediaWiki requires more construction and design on the part of the organizers, in addition to

maintaining appropriate funding levels. Zooniverse offers a platform and site builder that enables individualized project design and serves as a "temporary home for data generation" (Iranowska, 2019). With this background, one can begin to examine in detail the methodologies and successes of various crowdsourcing projects.

Independent Crowdsourcing Projects

Reese's 2016 article details the efforts of the University of Oklahoma Libraries to host a grant-funded transcription project of two Civil War diaries in conjunction with the sesquicentennial of the war. The first selection, the Garrett letters from the Charles Evans Collection, comprises 610 pages of letters between a soldier and his wife that were digitized specifically for the project. The second selection was the diary of Charles Kroff from the Sherry Marie Cress Collection, which had been digitized previously. The library hosted a basic website featuring a web-based transcription tool to facilitate the completion of the project, with the full site becoming available in August 2014. In just 81 days, 152 volunteers completed nearly 1,600 transcripts in order to completely reproduce the 787 manuscript pages (Reese, 2016). The organizers found that there were several above-average transcribers who represented the bulk of the volunteer effort. For the creators, the most time-consuming part of the project was transcription triangulation, or the method by which project staff validated the transcriptions and compiled the final text. This process was made more difficult by participants deviating from the established transcription guidelines. As a result of this project, both the completed transcriptions and high-resolution images of the source material are now available online.

Published in 2018, Paclíková et al. detail the results of a Czech-Bavarian crowdsourcing project to classify photographs held in various archives and reconstruct the historic appearances of border towns lost to war. PhotoStruk combines crowdsourced knowledge with "automated

image analysis and geo-referencing" in order to accurately determine the location of a photograph, make the information available to the public, and provide accurate metadata (Paclíková et al., 2018). Project organizers made photographs available through a specialized website and invited visitors to add comments or annotations, contribute location information, cite references, and add location names. The ultimate goal of the project is to create an archive as well as 3D models of towns based on historical evidence and to facilitate information exchange between the website and the archive. The project's website allows contributors to pinpoint the location from which a photograph was taken on a map and label known locations found in the archive's images. PhotoStruk also gave organizers the opportunity to examine how the public interacts with the photos and how interest can be generated in order to parse more of the data and recover lost information. By combining modern light detection and ranging (LiDAR) technology with historical aerial military photographs and community knowledge, it is possible to compare how human activity in the region has changed in the past century and beyond. This project represents a unique crowdsourcing approach, given its niche area of focus, both in the crowd that can assist with the project and the geographical region that benefits from it. Based on the author's explanation of the project, it is easy to see how the framework could be adapted to other parts of the world, given that enough relevant photographs are available for analysis.

Zooniverse-based Crowdsourcing Projects

Van Hyning's 2017 article details Zooniverse's origins, beginning with its first project, Galaxy Zoo, which was developed and managed by a team at the University of Oxford. Between 2007 and 2017, the site hosted over 100 projects, engaged over a million volunteers, and contributed to the generation of data for numerous scholarly articles (Van Hyning et al., 2017). As a member of the Oxford Zooniverse team, Van Hyning has unique insight into the 2016 grant

awarded to the organization by the Institute of Museum and Library Services (IMLS), which aimed to optimize the free, open-source platform for GLAM institutions. In particular, it was observed that GLAM institutions focus on crowdsourcing for collection accessibility rather than specific research questions. Online volunteering opens up opportunities for many more individuals to engage with an institution's collection than would ever be possible in person, with fewer concerns for the safety of the collection materials. Zooniverse facilitates the creation of various workflows through its Project Builder, which allows for the creation of sophisticated workflows while still being accessible to users of varying skill levels. Multiple volunteers independently classify each data point, and their responses are then compared to generate a consensus. Each project's organizers determine the number of repeated verifications. One of the unique features of Zooniverse is that the platform incorporates a Talk feature, or forum, which allows participants to discuss specific objects, ask questions, interact with researchers and volunteers, and save their own collections of material. Using the IMLS grant, the organizers of Zooniverse sought to optimize a workflow and determine the best practices for text and audio transcription. Additional goals included allowing data to be exported in formats that can easily be imported into GLAM databases, a scalable infrastructure, and a streamlined method for uploading data to the platform.

In 2018, Barber published a comprehensive examination of the positive impact of Zooniverse on solving the hidden collections problem within archives. The author defines hidden collections as any uncataloged or unprocessed materials that could benefit from collection or folder processing, or, in the case of digitized collections, increased searchability and description. Barber continues by identifying key crowdsourcing projects associated with libraries, including the Medici Archive Project and Smithsonian Transcription Center, both of which helped shape

how future projects would operate. Critical Zooniverse projects included AnnoTate and Shakespeare's World. In particular, Operation War Diary was chosen as a case study for a successful Zooniverse project. Operation War Diary represents a collaboration between the UK National Archives and the Imperial War Museum in an effort to transcribe roughly 1.5 million pages of British Army Regimental war diaries (Barber, 2018). Project organizers chose to label different fields using icons rather than MARC tags or Dublin Core elements in order to make the workflow accessible to volunteers without prior cataloging experience. Using a combination of free text fields and controlled vocabulary, five unique transcriptions were compiled for cross-referencing to form a consensus. Barber argues that projects such as these afford both the library and specific collections additional visibility, encouraging a wider community to engage with archival collections.

Lorrey et al.'s 2022 publication summarizes the experiences of the team behind the Zooniverse project Southern Weather Discovery. The project's main goal was to compile data collected by ships sailing in the New Zealand area between 1900 and 1950 in order to help evaluate the New Zealand Earth System Model for weather forecasting (Lorrey et al., 2022). According to the authors, "a total of 150,690 observations from 85 unique ships that embarked on 210 voyages were successfully captured by replicate keying from citizen scientists" (Lorrey et al., 2022). The term replicate keying refers to the number of repeated transcriptions required by the organizers before an item can be considered complete in the project's Zooniverse workflow. Initially, the project team required between ten and twenty transcriptions as they began the process of narrowing down the ideal number for maximum accuracy and efficiency. The authors felt that since data had to be collected from archives worldwide, the project presented a unique opportunity to draw attention to citizen science and the importance of archives and preservation.

To reduce fatigue and errors on the part of the volunteers, each page of a ship's log was broken down into separate workflows for ship position, temperature, and barometric pressure. As the project progressed, a decision was made to create separate boxes for each value rather than instructing transcribers to separate them within a single box using delimiters. Organizers found that some of the most time-consuming portions of the project included training personnel to select portions of logbook images for upload, create workflows for each unique logbook format, normalize numerical values, and respond to volunteer questions. The authors emphasize the importance of preparing enough data for the project to maintain momentum and encourage a greater knowledge-sharing practice between projects involving similar types of data.

Hawkins et al.'s 2023 article discusses the Rainfall Rescue project that the UK National Meteorological Archive launched in March 2020. The goal of this project was to transcribe the data present on Ten-Year rainfall dating back to the Seventeenth Century in the United Kingdom and Ireland. Given that these data sheets had a uniform design, the project consisted of two workflows, one that asked volunteers to transcribe the twelve monthly rainfall amounts and the annual total, with four replicate transcriptions per page. The second workflow focused on gathering location information, station reference, and grid number. Upon the project's completion, the organizers regretted not asking volunteers to include the data collector's name, which would have been helpful for cross-referencing. In a surprising turn of events, the transcription of the roughly 66,000 sheets was completed in just 16 days by around 16,000 participants (Hawkins et al., 2023). According to Hawkins et al., the transcribed information increases the number of observations available before 1961 tenfold and more than doubles the number of monthly observations available digitally in the United Kingdom. Contributors from the UK National Meteorological Archive noted that the raw data would not be beneficial for

long-term climate analysis since the observations often lack critical location information. In addition, gaps in the data exist due to the formation of the Republic of Ireland in 1949. Researchers found Zooniverse's Talk forums to be essential to fostering a community amongst volunteers and encouraging continued engagement. Since the completion of the transcription project, the data have been collated, quality-controlled, and attributed to any of the nearly 7,000 observation stations.

In 2024, two articles were published discussing the Maria Edgeworth Letters Project (MELP). Egenolf et al. (2024) provide a detailed overview of the project, explaining that by transcribing Edgeworth's letters, a digital open-access archive could be created to display raw text alongside encoded transcriptions in a searchable database. The project was launched with assistance from the NEH Advanced Institute in Digital Humanities to transcribe 744 letters comprising approximately 3,000 pages (Egenolf et al., 2024). Through the use of Zooniverse's Talk function, organizers were able to understand the most challenging parts of the transcription process, including Edgeworth's use of abbreviations, ampersands, and numbers. In order to ease the task of transcription, Zooniverse volunteers and research assistants from MELP created a guide for reading Romantic-era handwriting with a specific focus on Edgeworth. Another benefit of the Talk forums was that they allowed volunteer transcribers to ask questions regarding the content of the letters, thereby creating a greater understanding of the topics discussed within. Egenolf et al. felt that the project represented a perfect combination of the acquired expertise of volunteers, the scholarly expertise of editors, and the practical expertise of project research assistants. After transcription, MELP graduate assistants apply name authorities to the appropriate portions of each letter so they can be linked to Virtual International Authority Files.

Havens et al.'s 2024 article further details the technical aspects of the project, explaining that the goal of the transcription is not to correct unclear word choice or contradictory elements, but rather to preserve the content as it was originally written. The authors explain that the project draws content from over thirty libraries in North America and Europe, a fact that required extensive copyright negotiations. In order to facilitate transcription, organizers had to update and build upon various types of metadata and reduce copies of each digitized object to less than 1MB for upload to Zooniverse. Havens and collaborators also explain that crowdsourcing was preferred to AI use as artificial intelligence would eliminate the collaborative nature of the project and "remove the human from the humanities" (Havens et al., 2024). The authors elaborate on the contributions of volunteers, describing how the majority of the project's field guide on Zooniverse was either suggested or created by the volunteers. In addition, the Talk forums also allowed researchers and volunteers to collaborate to decipher difficult passages and discuss the historical context of the letters. Based on both of these publications, it is easy to see how Zooniverse facilitates unique interactions between researchers and volunteers that further the goals of each project and create useful insights for each project's continued development.

Another article published in 2024 by Teleti et al. summarizes the efforts and findings of Old Weather WW2, a Zooniverse project that digitized and transcribed data from United States Navy vessel logbooks of ships active during World War II. Researchers found that data from this time period was often lacking in detail or missing entirely since it had not been digitized. Due to the war, changes in methodology were poorly documented, leading to a general lack of understanding of the process and significant unexplained anomalies, particularly in water temperature values. The project's organizers focused specifically on ships that were active in the Pacific during World War II. According to Teleti et al., each ship's logbook recorded general

information such as the ship name or hull number, in addition to wind speed and direction, barometric pressure, air and water temperatures, visibility, and overall weather conditions, along with the ship's location. Weather data was recorded every hour, while the ship's coordinates were recorded three times a day. The project transcribed records from nineteen ships and separated the pages into workflows that transcribed navigation, barometric data, and temperature data separately. Records from 1943 and 1944 were part of a different Zooniverse workflow, as positional data was recorded in a separate logbook during the war. The organizers required three transcriptions per field for the object to be considered complete and decided the correct value based on consensus amongst the transcriptions. Overall, more than 630,000 unique records were transcribed for an average of 7,000 records per ship per year and 300 days' worth of records (Teleti et al., 2024). Organizers found that transcription tasks should be broken down into smaller segments to yield fewer errors and expressed interest in a function that would allow the pages to be transcribed in chronological order.

By analyzing projects hosted on both Zooniverse and independent websites, it is easy to see a general trend towards Zooniverse as the preferred host for such endeavors. Early organizers were forced to build their own websites in order to facilitate transcription or other such endeavors due to a lack of alternatives. Particularly in the years since the COVID-19 pandemic, Zooniverse's popularity and capabilities have expanded exponentially. With nearly three million registered volunteers, hosting a project on Zooniverse affords organizers access to a huge wealth of manpower and practical experience.

References

Alam, S. L., & Campbell, J. (2017). Temporal Motivations of Volunteers to Participate in Cultural Crowdsourcing Work. *Information Systems Research*, *28*(4), 744–759. Computers & Applied Sciences Complete. https://doi.org/10.1287/isre.2017.0719

Barber, S. T. (2018). The ZOONIVERSE is Expanding: Crowdsourced Solutions to the Hidden Collections Problem and the Rise of the Revolutionary Cataloging Interface. *Journal of Library Metadata*, *18*(2), 85–111. Academic Search Premier. https://doi.org/10.1080/19386389.2018.1489449

Egenolf, Egenolf, S., Havens, H., Richard, J., & Runia, R. (2024). Communities of Collaboration: Building the Maria Edgeworth Letters Project. *Studies in Romanticism.*, *63*(3), 335–368. https://doi.org/10.1353/srm.2024.a943147

Ellis, S. (2014). A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship. *Walter de Gruyter GmbH*, *64*(1), 1–10. https://doi.org/10.1515/libri-2014-0001

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. https://doi.org/10.1177/0165551512437638

Havens, H., Wilcox, E. A., Hale, M. L., & Kramer, J. (2024). From Archive to Database: Using Crowdsourcing, TEI, and Collaborative Labor to Construct the Maria Edgeworth Letters Project. *Digital Humanities Quarterly*, *18*(2). https://dhq.digitalhumanities.org/vol/18/2/000424/000424.html#recommendations

Hawkins, Hawkins, E., Burt, S., McCarthy, M., Murphy, C., Ross, C., Baldock, M., Brazier, J., Hersee, G., Huntley, J., Meats, R., O'Grady, J., Scrimgeour, I., & Silk, T. (2023). Millions of historical monthly rainfall observations taken in the UK and Ireland rescued by citizen scientists. *Geoscience Data Journal.*, *10*(2), 246–261. https://doi.org/10.1002/gdj3.157

Iranowska, & Iranowska, J. (2019). Greater good, empowerment and democratization? Affordances of the

crowdsourcing transcription projects. *Museum and Society.*, *17*(2), 210–228. https://doi.org/10.29311/mas.v17i2.2758

Lorrey, A. M., Pearce, P. R., Allan, R., Wilkinson, C., Woolley, J.-M., Judd, E., Mackay, S., Rawhat, S., Slivinski, L., Wilkinson, S., Hawkins, E., Quesnel, P., & Compo, G. P. (2022). Meteorological data rescue: Citizen science lessons learned from Southern Weather Discovery. *Patterns.*, *3*(6), 100495. https://doi.org/10.1016/j.patter.2022.100495

Paclíková, K., Weinfurtner, A., Vohnoutov, M., Dorner, W., Fesl, J., Preusz, M., Dostálek, L., & Horníčková, K. (2018). Geoinformatics and Crowdsourcing in Cultural Heritage: A Tool for Managing Historical Archives. *Agris On-Line Papers in Economics & Informatics*, *10*(2), 73–83. Academic Search Premier. https://doi.org/10.7160/aol.2018.100207

Reese, J. S. (2016). Transcribing the Past: Crowdsourcing Transcription of Civil War Manuscripts. *Archival Issues: Journal of the Midwest Archives Conference*, *37*(2), 59–74. Academic Search Premier. https://doi.org/10.31274/archivalissues.11019

Van Hyning, Hyning, V. V., Blickhan, S., Trouille, L., & Lintott, C. (2017). Transforming Libraries and Archives through Crowdsourcing. *D-Lib Magazine*, *23*(5–6). https://doi.org/10.1045/may2017-vanhyning