https://doi.org/10.2352/issn.2168-3204.2019.1.0.33

Crowdsourcing the Smithsonian: Developing and Maintaining the Smithsonian Transcription Center and the Digital "Volunpeer" Community

Caitlin Haynes[1,2], Janet B. Abrams[2], and Michael Schall[1]; [1]Quotient, Inc., Columbia, Maryland, and [2]Smithsonian Institution, Washington, D.C., United States

Abstract

The Smithsonian Institution has digitized millions of its archival, library, and museum collections over the past decade, but this work is only the first step in providing and improving collection accessibility and use. To further increase the discoverability and engagement of their digitized materials, the Smithsonian developed the Smithsonian Transcription Center—a pan-institutional crowdsourcing project that enlists the public (anywhere in the world) in transcribing and reviewing digitized field notes, diaries, letters, specimen catalog cards, and more. This paper and presentation will discuss the development, growth, and achievements of the Transcription Center, along with the ways in which the lessons learned by Smithsonian staff can inform others interested in creating their own crowdsourcing platforms.

Introduction

The Smithsonian Institution strives to preserve cultural heritage, discover new knowledge, and share its resources with the world. As technology has advanced, digitization of the vast library, archival, and museum collections held within the Smithsonian's nineteen museums and research centers has become a central aspect of fulfilling the Institution's historic mission of increasing and diffusing knowledge. Yet, as success in digitization continues, questions and challenges emerge about how best to proceed. How can we ensure that digitized content is being made available for study and use by the public? How can staff from different parts of the Institution work with each other and with volunteers to boost awareness and use of Smithsonian collections by researchers and students? If collections include handwritten materials, is it enough to provide the public images of the pages? Could more be done to unpack the content?

These questions motivated the development and growth of the Smithsonian Transcription Center (TC) in 2013. Designed as a publicly accessible website (https://transcription.si.edu/) where digitized collections from across the Smithsonian are hosted and can be transcribed and reviewed by the public, the Transcription Center serves to increase access to our diverse collections, engage the public in creating new pathways of learning and knowledge creation, and strengthen Smithsonian engagement with researchers and other communities of interest. Over the past five years, the Transcription Center development and support team of the Office of the Chief Information Officer ("TC team") has worked collaboratively with colleagues from Smithsonian libraries, archives, and museums, as well as thousands of digital volunteers participating in transcription, to expand and continuously assess and improve the platform.

Not Just Another Crowdsourcing Site: The Smithsonian Transcription Center's Design and Growth

First and foremost, the Smithsonian Transcription Center was designed as a flexible platform for collaboratively engaging the public in improving digitized Smithsonian collections. The TC was created specifically as a site that allows volunteers to work on – and move between – multiple different projects at their own pace, and collaborate and communicate with other volunteers and Smithsonian staff through multiple avenues. It also provides a service for Smithsonian staff by offering a flexible and integrated backend. This allows participating museum units (to date, departments and units within thirteen of the nineteen Smithsonian museums and various research centers have participated in the TC) around the Institution to easily import and export their digitized materials into the TC, edit, review, and approve projects, communicate with volunteers, and monitor progress and engagement. Technically developed utilizing a Drupal module and web application, the site is integrated with existing Smithsonian collection and information management systems and thus accommodates the importing of new projects with many different material format types. For example, image assets (used to create the pages of individual Transcription Center projects) can be imported from the Institution's digital asset management system (DAMS), the image delivery service (IDS), using unique encoded archival description (EAD) identifiers from ArchivesSpace records, or from unique metadata in catalog records of the enterprise digital asset network (EDAN) platform. Completed TC projects are conversely always connected back to these internal systems. Once transcriptions are reviewed and marked complete, the transcribed text is automatically indexed and connected to the original item's catalog or collection record in the Smithsonian's central online databases, Collections Search Center and the Smithsonian's Online Virtual Archive (SOVA). Transcription Center work by digital volunteers thus becomes text-searchable and more easily discoverable and accessible.

Beyond simply creating readable transcriptions of digitized content, however, the Transcription Center is also utilized as a way to create and enhance new catalog records and collection metadata for Smithsonian materials. While millions of the Institution's collection items have been processed, catalogued, and digitized, the sheer volume of materials, along with staff and funding constraints, means many collections remain uncatalogued, inadequately described, and inaccessible to remote researchers. Enlisting the help of the Transcription Center's volunteer community to transcribe collection information offers an efficient solution. In 2014, thousands of newly digitized bumblebee specimens

and labels from the National Museum of Natural History's (NMNH) Department of Entomology were imported as projects into the TC where digital volunteers were asked to transcribe the specimen labels into a predesigned template [1]. These transcriptions were then used to create text-searchable catalog records in the Smithsonian's online database Collections Search Center. The Bumblebee Project, as it was aptly named, resulted in the transcription of data from over 44,000 bumblebee specimens, increasing accessibility for researchers around the world and revealing new information on change over time in pollinator populations. Since then, other Smithsonian museum departments – including the NMNH Department of Botany and the National Museum of American History's Archive Center—have partnered with the TC's digital volunteers to transcribe voluminous collection data, creating and enhancing catalog records and revealing new research along the way.

The flexible platform and integrated-web infrastructure contribute significantly to the success of the Transcription Center as a crowdsourcing platform. Volunteers routinely acknowledge how much they like the ability to jump around between various projects with no defined time or progress commitment. Whether for an entire day or simply ten minutes, digital volunteers can transcribe and review scientific field books, historical diaries, museum administrative records, or all of the above from the comfort of their own home. This system also works well because it allows staff from across the nineteen Smithsonian Institution museums to host and manage their digitized materials, and interact with volunteers and other colleagues in one centralized platform. While the Smithsonian has other centralized information systems and databases (as mentioned in the previous paragraphs), the Transcription Center is the only current site hosted by the Institution that provides an interactive platform with content from all systems, units, and departments within the complex organization of Smithsonian museums, libraries, and archives. To date, more than 12,000 volunteers have collaboratively transcribed and reviewed almost 450,000 pages of field notes, diaries, ledgers, logbooks, currency proof sheets, photo albums, correspondence, biodiversity specimen labels, and much more from Smithsonian collections.

Prioritizing "Volunpeers": The Importance of Collaboration and Communication

The technical design of the Transcription Center as an integrative, flexible system, encourages participation from the public and internal staff, and played a critical role in the platform's growth over the last five years. Yet equally important to the TC's continued success is the emphasis placed on communication and collaboration with digital volunteers. Simply building an interactive platform for the public to engage with is not enough to establish or sustain a volunteer community. Technical management and coordinated public outreach must be given the same staff attention for this kind of project to succeed. Early on, the TC team learned that multiple forms of general and targeted outreach with volunteers was essential to encourage and build public engagement, improve the site, and learn more about our collections. Rather than recruiting volunteers through a single call to action, the TC continuously invites public participation and encourages volunteer growth through ongoing communication.

Digital volunteers participating in TC work alongside staff from the TC team and participating Smithsonian museums to share knowledge and improve collections. This emphasis on collaboration undergirds all engagement projects and campaigns, and led to the creation of the term "volunpeer." Coined shortly after the creation of the Transcription Center itself, the word volunpeer refers to our digital community in a way that more appropriately reflects their role as equal peers and partners in this project. teaching us at the Smithsonian just as much as we are sharing with them. Because of this, the TC team has included a dedicated project coordinator since 2013. The coordinator is tasked specifically with guiding and responding to volunpeers through email communication, social media, and other outreach, and working directly with Smithsonian staff contributing materials to the platform to post their projects and monitor transcription progress.

Through the first few years of the TC, the project coordinator collaborated with contributing staff and colleagues within the Smithsonian's Office of the Chief Information Officer (OCIO) to plan and implement multiple structured outreach efforts and everyday engagement practices. A monthly electronic newsletter was designed through MailChimp to share updates, achievements, and highlights to volunteers, instructional videos and behind-the-scenes Facebook Live or Google Hangout videos were planned and created, and outreach strategies for multiple social media platforms were developed. Public engagement proved most successful when outreach was approached through a combination of specific campaigns alongside ongoing promotion of projects by the TC team and participating Smithsonian staff. While volunpeer progress, new projects, and transcription tips were communicated regularly on social media, the TC homepage, and digital newsletters, deeper explorations of featured collections, staff and volunteer spotlights, and behind-the-scenes work, were shared in targeted campaigns only periodically. Special blog posts, social media blasts, and outreach videos prepared by participating staff and the TC coordinator, invited volunteers to go beyond ongoing updates and learn even more about new projects. The adoption of this two-pronged outreach approach offered the increasingly diverse community of digital volunteers multiple ways to dive into Smithsonian collections, ensuring continued public interest and engagement [2].

TC staff also designed into the platform and outreach plans multiple avenues for digital volunteers to communicate with each other and Smithsonian employees. Not only was a resource email account created (transcribe@si.edu) for volunteers to email with the project coordinator directly in regards to project questions, feedback, and discoveries, but feedback contact form entries were also added to the platform – appearing on every page for volunteers to use. Individual project pages within the Transcription Center also include a "notes" box underneath the text box for transcriptions. This area was designed as a place for volunteers to leave questions, comments, and additional information for each other and for internal staff, and conversely, for internal staff to respond to inquiries and further explain any editing or instructions. On the administrative end, Smithsonian staff from participating museums can retrieve statistics related to their unit's projects and the volunteers working on them. This not only allows for participating museum units to report their progress with the Transcription Center, but also provides further opportunities

to reach out to volunteers personally to encourage more participation, express gratitude for project progress, and share information.

Additionally, volunteers can – and most often do – communicate through TC's social media platforms (Facebook, Instagram, Twitter). Twitter is particularly used, with volunteers using the #volunpeer to tweet their discoveries, tips and tricks for challenging projects, welcome newcomers to the project, and seek help from others. While transcribing U.S. naturalist Vernon Bailey's field books from an 1890 expedition, volunpeers in 2017 were caught by surprise when they read in his field book that he had captured and eaten a golden eagle. A lively discussion was sparked among volunpeers about this on the TC Twitter account, with staff from the Smithsonian Institution Archives and the National Museum of Natural History (NMNH) chiming in about the lack of food available in the nineteenth century (particularly on expeditions), and the current location of the eagle in question as a specimen in the NMNH Department of Birds [3]. TC volunteers have also used social media to share personal connections to information found within Smithsonian materials and provide further background on species names or historical topics, enriching collections beyond transcription. This continuous communication between digital volunteers and Smithsonian staff is central to the success of the TC. Volunteers are not simply transcribing featured projects, they are actively working alongside the Institution to improve collections. As this partnership has developed over time, the caliber of work produced by the volunteer community has continued to improve, with questions or fears of quality and accuracy diminishing rapidly. Volunteers are deeply motivated, dedicated, and engaged and express personal pride and ownership over their TC achievements.

Moving Forward: New Technologies and Platform Development

In addition to ensuring efficient operation of the Transcription Center platform and striving to expand and motivate our volunteer corps, the TC team works to improve upon and evolve the original system, incorporating new technologies both to take advantage of emerging opportunities and to address feedback received from participating staff and volunteers.

The platform's success to-date is due in great part to focused attention by staff to what works best for the volunpeers. For example, it was observed that transcription-and-review of some very large projects (e.g., many book pages or specimen collection records totaling more than 1,000 pages, etc.) took longer on a per-unit basis than transcription of smaller projects. The TC team divided the large projects into smaller lots, and transcription pace improved. In addition, the team observed that volunteer productivity was higher when a variety of materials across Smithsonian holdings (art, history, culture, science) was offered than when choices were limited. Now the team strives to offer visitors to the site a broad selection of collections to work on, a selection that represents the tremendous breadth of the Institution. Currently, the TC offers projects on the following themes and topics: African American history, women's history, American experience, art and design, biodiverse planet, the Civil War era, the Freedmen's Bureau, mysteries of the universe, the Smithsonian's Field Book Project (a collaborative initiative between the Smithsonian Libraries and the Smithsonian Institution Archives to digitize historical field books), Native American and Indigenous history, and world cultures.

The TC team has also continued to evolve public outreach and engagement strategies in response to user feedback. Online public participation in social media campaigns promoting new projects and exploring featured collections has grown to over 9,000 followers across multiple platforms. Updates, volunteer achievements and discoveries, and background information are still shared regularly, and volunpeers continue to actively communicate with each other and Smithsonian staff via Twitter and other avenues. New outreach campaigns, resources, and engagement tools have also been created to expand the reach of the TC and further share Smithsonian collections. Once such campaign, a monthly twitter chat known as #TCImpact, was launched in 2018 in response to inquiries from Smithsonian staff and volunteers about the use of transcriptions in research projects. One day each month, the TC Coordinator hosts a dialogue on Twitter between volunteers, researchers, and internal colleagues to share various data, analysis, and research made possible by the text-searchable, transcribed content created in the TC. #TCImpact chats revealed the far-reaching influence of completed transcription projects, from Smithsonian Gardens' staff using transcribed letters from the Burpee Seed Company Collection to create digital maps of vegetable crops across the United States [4], to the National Museum of African American Culture and History utilizing the more than 20,000 pages (and counting!) of transcribed Freedmen's Bureau Records to uncover genealogical information and improve understandings of the Civil War era and post-emancipation life [5]. The online community of volunteers also shared personal motivations for participating in transcription during #TCImpact chats, with many noting that TC projects provided a much-needed alternative to volunteer tasks that required physical activity. Transcription allowed them to contribute something of value, be part of a community, and take an active role in Smithsonian projects from the comfort of their own homes [5].

Colleagues, educators, and other users have also requested the creation of targeted educational resources utilizing the TC platform and the historical information unlocked through transcription of Smithsonian collections. In response to this, the TC team partnered with the Smithsonian Center for Digital Learning and Access (SCLDA) and other Smithsonian units participating in TC projects, to curate Transcription Center focused digital collections and educational activities through the Smithsonian Learning Lab. This free, interactive platform provides educators, students, and others the ability to discover millions of authentic digital resources, creating content with online tools, and share in the Smithsonian's expansive community of knowledge and learning. Currently, two Learning Lab collections specifically related to the 2019 National History Day theme of Triumph and Tragedy, have been authored by the TC team: "Exploring WWI through Transcription" and "U.S. Reconstruction: 1865-1877" [7]. These collections bring together digitized images, transcriptions, and guiding questions and tips for analysis for students and educators to use as a launching point for further research. Through additional internal collaborations, the TC team is currently planning the creation of additional Transcription Center projects aimed specifically at K-12 educators and students, including a dedicated resource page

within the TC website, more Learning Lab collections, and instructional videos.

Continued feedback from volunteers and Smithsonian colleagues also influences technological developments within the Transcription Center. As the TC's success has grown, both internal and external users have expressed interest in investigating new ways that the TC platform and community can be leveraged to promote and explore a wider array of Smithsonian collections and further increase accessibility. Currently, transcription of audio and video content is a major development focus. Features to be released in the near term include the ability to segment long audio files into manageable portions, so that volunteers can work collaboratively and simultaneously transcribing different parts of the same project. To date there are over 3,600 sound recordings from libraries, archives, and museums at least partially digitized and online in the Smithsonian Collections Search Center database, and over 114,000 audio materials that have yet to be digitized. Many of the audio recordings currently available online lack transcriptions and/or captioning, making them inaccessible to individuals who are deaf and hard-of-hearing. Moreover, new accessibility compliance regulations prevent these materials from being made public on Smithsonian websites without transcriptions or captions. Harnessing the power of the TC's volunteer community to transcribe and review these digitized sound recordings will provide a cost-effective solution, by allowing staff to generate closed captioning files for audio and video from the transcriptions, while also engaging the public and revealing new content within Smithsonian audio-visual collections.

Another recent technical innovation that will benefit research and scholarship at the Smithsonian and around the world is the integration of IIIF (International Image Interoperability Framework) [8]. During the summer of 2018, the Smithsonian launched support for IIIF for an initial dataset of 800,000 connected digitized assets. Over time more assets will become IIIF-aware and will include connected data, such as transcriptions from the TC, with images being delivered via IIIF Presentation manifests. While still in the preliminary stages, the implementation of IIIF at the Smithsonian ensures that digitized content – and corresponding transcriptions produced by TC volunpeers – will be discoverable and accessible to researchers around the world on an unprecedented level.

Conclusion

Since 2013, the Smithsonian Transcription Center has successfully engaged a growing community of digital volunteers in improving digitized collections from around the Institution. More than 12,000 volunteers collaborate with each other and participating Smithsonian staff to transcribe, review, and explore historical materials. To date, nearly 450,000 pages of diaries, correspondence, specimen labels, and more, have been completed. This success is a direct result of the time and attention the TC team has devoted to building and maintaining the volunpeer community. By continuously collaborating with digital volunteers and colleagues, TC staff have been able to create and sustain an evolving, interactive platform. Digital volunteers have been instrumental in helping guide platform and technological enhancements, project choices, and outreach efforts. Staff have learned how best to improve upon past success, implement new kinds of projects, share information, and garner interest. As the Smithsonian Transcription Center grows and develops, we look forward to learning new tactics and ideas from our colleagues and volunteers, and similarly hope that TC approaches and strategies can help inform other organizations as they build crowdsourcing projects and "volunpeer" communities.

References

[1] The bumblebee specimens held by the National Museum of Natural History (NMNH) were digitized as part of a rapid capture conveyor belt digitization project led by the Smithsonian's Digitization Program Office. To learn more about this, see the presentation from NMNH here: https://www.idigbio.org/wiki/images/8/88/Kimberley_NMNH_RapidCapture.pdf; and the Bumblebee Project instruction page on the Transcription Center website: https://transcription.si.edu/instructions-entomology.

[2] Lesley Parilla and Meghan Ferriter (2016) Social Media and Crowdsourced Transcription of Historical Materials at the Smithsonian Institution: Methods for Strengthening Community Engagement and Its Tie to Transcription Output. The American Archivist: Fall/Winter 2016, Vol. 79, No. 2, pp. 438-460.

[3] To read more about volunteer collaborations and communications via Twitter, including additional details on the Golden Eagle discovery visit the Smithsonian Collections Blog: http://si-siris.blogspot.com/2017/10/uncovering-history-with-smithsonian.html..E. Jones, An Inexpensive Micro-Goniophotometry You Can Build, Proc. PICS, pg. 179. (1998).

[4] AAG-Burpee-Sample Garden Vegetation, 1924 Burpee Contest Letter - Transcription Center, April 2015, Interactive Map, https://www.google.com/maps/d/u/0/viewer?mid=12prcplVWKedvw3kYdgzy7FBdSXs&ll=41.94246006395072%2C-93.62590845&z=4.

[5] The Freedmen's Bureau Records at the National Museum of African American History and Culture, Smithsonian Institution, https://nmaahc.si.edu/explore/initiatives/freedmens-bureau-records.

[6] To see all #TCImpact chats, visit https://twitter.com/search?q=%23TCImpact&src=typd.

[7] Smithsonian Transcription Center Learning Lab Collections, 2018, https://learninglab.si.edu/profile/smithsoniantranscriptioncenter.

[8] To learn more about IIIF integration at the Smithsonian Institution, visit https://iiif.si.edu/.

Author Biography

Caitlin Haynes is the Coordinator for the Smithsonian Transcription Center. In this position, she is responsible for digital volunteer engagement, outreach, and the coordination of new projects for crowdsourced transcription. She holds an MA in United States History and an MLIS in Archives and Records Management from the University of Maryland, College Park. Prior to serving in her current position, Caitlin was the Reference Archivist at the National Anthropological Archives at the Smithsonian's National Museum of Natural History from 2015-2018, and served as a Research Associate for the Papers of Abraham Lincoln from 2013-2015