White Paper for Enabling and Reusing Multilingual Citizen Contributions in the Archival Record

National Endowment for the Humanities Grant HAA-269051-20

Allyssa Guzman, Albert A. Palacios, Ryan Sullivant
November 30, 2022

Project Summary

This project arose from the growing consensus that representation has not been enough to diversify the digital cultural record. Rather, as digital humanists, archivists, and librarians have pointed out, representation without the participation of non-Anglophone and minority groups has recreated historical exclusions, which now stem from a lack of technological or multilingual resources that facilitate access and engagement with materials in other languages (Priani Saisó et al. 6; Caswell et al.; Bow and Hepworth; Risam). The message in these critiques is resounding: the field needs to promote and support cultural and linguistic diversity. Part 1 of this project provided an avenue for non-English literate communities to meaningfully engage and contribute to the Digital Humanities through the interface internationalization and translation of an open source digital scholarship platform–FromThePage. Part 2 of this project enhanced FromThePage's collection management capabilities and exports to facilitate the development of workflows for preserving and reusing collaborative scholarship.

FromThePage (FtP) is a tool for the collaborative transcription of primary source materials, particularly those that are handwritten or not readable by optical character recognition (OCR) technologies. Additional features allow for the translation of texts, indexing and extraction of key subject terms (e.g. people and place names), version control, and tracking statistics on collaborator contributions and attribution, among others. Prior to our work, FtP was only accessible to those who can read English. This project internationalized FtP through 1.) the reconfiguration of the platform's infrastructure to enable a multilingual interface, and 2.) the subsequent translation of the tool and its user guides into Spanish and Portuguese. This was a joint effort between the University of Texas Libraries (UTL) and Brumfield Labs, the creators of FromThePage.

Project Goals

Our project had two principal goals. First, we wanted to break down barriers to access that Anglophone-centric technology stacks create for our Spanish and Portuguese language digital collections. Specifically, we wanted to address this barrier in FromThePage (FtP), a Brumfield Labs digital humanities tool we use to enable the collaborative transcription, translation, and indexing of primary source materials at UTL. By internationalizing the platform, FtP went from being an English-only tool to a platform that not only had a Spanish and Portuguese interface, but also the flexibility to incorporate other languages.

Second, we wanted to develop workflows and recommendations for preserving and reusing collaborative scholarship in digital archival records. We believe that every individual who meaningfully contributes to the interpretation and understanding of our collections should be given credit for their efforts in the archival record and in scholarship. Part of this work necessitated enhancements of FromThePage's collection management capabilities and exports to facilitate the movement of primary source materials into FtP and contributed scholarship back into digital assessment management systems.

Principle Outcomes

In the first year of the project, we produced significant updates to the FtP transcription software and introduced new features to allow for better management of collections within FtP, and to better facilitate the transcription and translation of large collections of hand-written documents. We wrote introductory FtP user guides first in English, and then translated them successfully into Portuguese and Spanish, providing an avenue for non-English speaking individuals and communities to engage with the digital tool. We also developed course curriculum materials including the workshop presentation slides for presentations held throughout the year. To support Brumfield Labs in the development of additional software features, we conducted a user needs assessment of FtP project owners. The results of this study were collected into a report that was shared with Brumfield Labs and the FtP user community.

In the second year of the project, we added additional features to the FtP software that will better allow project managers to reuse the results of transcription and translation projects in other software systems, enabling greater discovery via full-text searching and opening up cultural heritage materials for use in research. We also drafted best practices around both attribution of project contributions and reuse of transcriptions. In order to expand on this work, we hosted a virtually-held symposium, *After the*

Transcription, to bring project managers working on collaborative transcription projects into a community of practice. In addition to our work on reuse, we also worked with community partners in Mexico and Brazil to vet the translations of the FtP interface into Portuguese and Spanish.

Plan of Work

Metadata Import

Between February and May, 2020, Brumfield Labs improved browsing and discoverability within FtP by developing code to enable the import and display of descriptive metadata. Our Metadata team provided real metadata examples from UTL collections to Brumfield Labs for testing. Once imported, the collection owner can designate which fields they would like to use for faceted browsing. The metadata import is standard-agnostic and allows users to upload metadata from any standard. The updated FtP source code was made available on GitHub.

User Needs Assessment

From February to September 2020, we developed and distributed an IRB-approved survey among FtP collection owners to determine which export functions users most needed to maximize the reusability of citizen contributions to their FtP collections. Follow-up interviews were then conducted after initial survey responses were collected, and results were coded and analyzed for trends by Joshua Ortiz Baco. A detailed needs assessment report was published in Texas ScholarWorks, UT Libraries' institutional repository. It includes a list of desired export formats, as well as information about project owners' use cases for FtP, including considerations around crediting crowdsourcing participants. We learned that many of the project owners are interested in crediting participants, but that they lack the time and resources necessary to do so.

Internationalization and Translation

The internationalization and translation of the interface into Spanish and Portuguese was undertaken from May to September 2020. Brumfield Labs extracted the text from the FtP interface elements (i.e. navigational tabs, button labels, text entry descriptors, etc.) and assigned each element a unique key. Certified translator Joshua Ortiz Baco then translated the element dictionary into Spanish and Portuguese. Brumfield Labs then added the dictionary translations to the code so that they programmatically feed into the interface's elements when a user selects the platform's language, and updated the UTL instance for user testing. The restructured FtP source code including features

to switch between Spanish, Portuguese, and English has been published on GitHub and is available on FromThePage.com and the UTL's FtP installation. As we are committed to maintaining the Spanish and Portuguese interface, we will continue to translate user-facing text in new features and functionality as the need arises.

Bryce Mclin and Albert A. Palacios translated the FtP user guides into Spanish and Portuguese. The updated version of FtP was distributed to our community partners to test the localized versions of the software and user guides between January and August, 2021.

Albert A. Palacios created an FtP collection of materials from the Fondo Real de Cholula, held by the Archivo Judicial del Estado de Puebla, located in Puebla, Mexico. Dr. Lidia Gómez García documented her students' questions and feedback throughout the semester and provided it to the Project Co-Managers and Bryce Mclin, who addressed translation and functionality issues. These corrections were implemented in the FtP codebase in the September 2021 and February 2022 releases. At the conclusion of the semester, Dr. Gómez García corrected and completed final transcriptions in preparation for export and ingestion of these into UTL's digital repositories.

Albert A. Palacios created an FtP collection of materials on the Eldorado quilombo from the Articulation and Advisory Team to Rural Black Communities (EAACONE) digital holdings for use by the community. He worked virtually with EAACONE to observe platform use, collect feedback on the Portuguese translation of the interface and user guides, troubleshoot any issues, and provide additional platform training.

Additional Export Options

From February 2021 to September 2021, Brumfield Labs added the additional export features indicated by the user needs assessment. This work was originally planned for the previous year, but due to concerns around Ortiz Baco's availability, we decided to focus on his translation work in 2020 instead. The additional export options added include various document formats, a static site, an export to Voyant Tools, and the option for collection owners to allow volunteers the ability to download their work.

Document Export

Users can now download transcribed documents for reuse in plaintext, Word docx, PDF and HTML. There are two PDF export options, one which includes the transcribed text and a "Facing PDF," which includes images and transcribed text on facing pages.

Static Site Export

Exporting work to a static site using the static site generator Jekyll provides a structured set of HTML pages that can be viewed on any web browser without dependencies beyond a simple web browser. For instance, sites can be published to Github pages or even be shared on a thumb drive. These internally consistent static sites are also ideal for digital preservation.

Export to Voyant Tools

While FromThePage has been used in introductory Digital Humanities Methods classes, reuse of the transcribed text in analytical tools has been a challenge. A new “Analyze in Voyant” feature provides students and scholars with a simple system for analyzing their own texts in Voyant Tools, leveraging Voyant’s API and the new FromThePage export formats.

Volunteer Export UI

Providing volunteers with access to their own work is a best practice in crowdsourcing ethics. In order to give volunteers access to their own work, project owners now have the ability to turn on a user interface with full access to these export formats.

Work Toward Best Practices

From January 2021 through July 2022, we worked with a group of metadata experts and repository staff to establish best practices for preservation, access, and reuse of transcriptions as well as determining the proper attribution of citizen contributions in the archival record and scholarship. Our group engaged with issues of attribution, rights, and planning for preservation, reuse, and improving accessibility. The repository landscape at UTL is complex and includes multiple different technology stacks as well as a variety of content types from digitized primary source materials to scholarly gray literature and datasets. Because of this, we concluded that there would be no single workflow or solution that would allow round-trip integration from importing materials into the originating repository into FromThePage and incorporating the transcriptions, translations, and contributor attributions back into the repository that would apply to all of our systems.

The attributions were a particular challenge. As previously mentioned, we do not consider contributors as simply a “sourced crowd;” they are often experts in the subject matter represented in these primary sources, and their collective scholarship (i.e. transcriptions, translations, indexes) is worthy of preservation, therefore we focused on solutions that would allow us to credit contributors. Ultimately we concluded that for many of our systems, we would treat the transcriptions or translations as new objects

with their own unique descriptive metadata, which would allow us more freedom to credit contributors. For the majority of our projects, we are depositing transcriptions and translations in the Texas Data Repository to take advantage of the robust descriptive metadata options and to promote downloading and reuse of the materials.

When our anticipated travel fell through due to the global pandemic, we decided to expand this work to include organizing and hosting a symposium of invited participants working in crowdsourcing and community transcription, which was held in April 2022 over Zoom. A keynote presentation by Dr. Mia Ridge (the British Library), Dr. Megan Ferriter (the Library of Congress), and Dr. Sam Blickhan (Zooniverse) outlined current challenges and opportunities in the field of collaborative transcription. Conversations were convened around the definition of a "meaningful" contribution, rights over transcriptions, crediting of student and citizen collaborators, and the incorporation of contributed work into institutional repositories. Participants presented case studies from their work and then were guided in group discussions to further expand our collective knowledge of these topics.

Results

Since we released the Spanish and Portuguese versions of FtP and accompanying user guides, we have seen an increase in the platform's use by communities using languages other than English. For example, community scholars and history students from Puebla, Mexico, have used the tool to work on a project and practice their Spanish paleography skills. Besides using the transcription functionality, some students have been using the indexing/encoding feature to start identifying information for future extraction and analysis. We believe this work is demonstrating how the platform is enabling a broader community's meaningful engagement with the Digital Humanities, one of our central goals.

As part of a "Spanish Paleography & Digital Humanities Institute" sponsored by a separate NEH-AHRC grant, 60 colonial Latin Americanists—including 35 graduate students, 8 junior faculty, 8 tenured professors, 5 archive/library professionals, and 4 independent researchers from 11 countries and 18 U.S. states—primarily used the Spanish version of the platform and guides to transcribe 45 documents (approximately 540 pages) in our collections. Some participants expressed that the tool was user friendly and did not identify incorrect or missing translations, suggesting that the grant work was successful.

We have also started to conceptualize and implement ways to attribute citizen contributions in the archival record as we ingested transcriptions into our repositories.

After numerous meetings with our project metadata team, we decided to treat the transcription and scans of the original archival source as separate digital objects with their own digital records to preserve the provenance of their creation, using the MODS Related Resource field to link the records. Ultimately, we decided to deposit the transcriptions in our data repository and we developed a workflow and descriptive practice for the assets we ingested into the Texas Data Repository, which is a Dataverse platform.

Thanks to the workflows we developed, we started depositing and publishing citizen transcriptions in our LLILAS Benson data repository (https://dataverse.tdl.org/dataverse/blac). Each deposit is fully described in Spanish and English, gives due credit to all who contributed to the creation of the digital texts, and provides open access to page-level transcriptions in TXT format for reusability (example: https://doi.org/10.18738/T8/W3WHZG). Since the platform auto-generates a DOI of the deposited text, we have been able to send this out to the students and scholars so that they can include it in their curriculum vitae as a publication. This not only provides evidence of their academic work, which is often required by their academic institutions, but also gives them an incentive to continue their participation in the collaborative research project. A good number of these scholars have asked for guidance on how to include these publications in their CV and others have also inquired about reusing these transcriptions in their research and/or digital humanities projects. This demonstrates the impact this work has had in the professional development of scholars in the discipline.

The grant project is also contributing to information and archival studies and profession. The user study and needs assessment we completed in April 2020 yielded an environmental scan that we were able to share with other coordinators of collaborative scholarship projects using FormThePage. These coordinators were either scholars or professional staff in cultural institutions. The report demonstrated that project coordinators are just beginning to consider ways to credit citizen contributors ethically. However, some potential initial best practices emerged from that report that we were able to reflect on, build on, and share out. Since then, there has been some interest in the FromThePage community to continue the conversation with the hope that we can flesh out best practices.

The platform's translation is also facilitating an unexpected collaboration that is furthering other information studies work. Using the Spanish interface, Dr. Isabel Galina has been coordinating an effort that seeks to create an authorized Spanish edition of Trevor Owen's The Theory and Craft of Digital Preservation (John Hopkins University Press, 2018) for digital preservation professionals in Latin America. Although contributors to that project understand English, having the option to change the platform's language to their language appears to make the platform more approachable.

Changes and Problems

Due to the economic and health repercussions of the global pandemic beginning in March 2020, our project partner in Brazil, Equipe de Articulação e Assessorias às Comunidades Negras do Vale do Ribeira (EAACONE), could no longer dedicate a staff member to the project without financial support. Since the pandemic also derailed the project's travel plans and staffing needs changed, we decided to reallocate some of the grant funds to add another consultant fee budget line to support our Brazilian community partner. These funds enabled EAACONE to assign two staff members with language expertise to review the Portuguese translations (platform and user guides) we produced. Resulting financial mitigations and turnover at the university's accounting office also led to delay in processing contracts for our project consultants. As expected, this also delayed the start date for their work.

Numerous project expenses did not materialize due to the pandemic. These included travel and expenses associated with community training workshops in Mexico and Brazil we had to cancel, such as the project laptop and catering. Since the university procured an institutional license for Zoom and the user study we conducted did not require us to purchase a Dedoose license, we also had these unspent funds. We were able to reallocate funds to add another consultant fee budget line to support our Brazilian community partner as they reviewed the Portuguese translations (platform and user guides) we produced and created transcriptions for us to pilot ingestion workflows. Second, we were able to increase the consultant fee for our Mexico partner to increase their production of transcriptions for us to test workflows.

As the project team was unable to travel during this project, we redirected our budgeted funds for travel to putting on a virtual symposium entitled "After the Transcription," in order to expand on our work to conceive workflows and recommended practices for the preservation, long-term access, and attribution of collaborative transcription work. The symposium was primarily meant to be an extended working group session for cultural repository professionals and scholars who are actively managing collaborative ("crowdsourced") transcription projects to discuss approaches for incorporating and crediting contributions in institutional repositories.

Conclusion

Lessons Learned

The user study we conducted early on demonstrated that, generally speaking, collaborative transcription project coordinators using the FtP platform had not considered the attribution of contributed transcriptions. However, they were interested in crediting the participants in their crowdsourcing projects for their work. The main reason

why they had not considered it was because they were primarily focusing on getting their project “off the ground”, actively cultivating interest and a transcribing community around their collections. Since many of them were in the beginning phases of their “crowdsourcing” projects, most did not have enough completed transcriptions to start thinking about their subsequent export, much less ingestion into other access and preservation platforms. For most of the interviewees, the study prompted them to start thinking about the eventual exportation of transcriptions and translations from their projects for reuse in other systems.

As we conceived potential preservation workflows, we also came to the conclusion that collaborative transcriptions require their own archival record. In our attempt to incorporate the provenance of digital texts into the original manuscript’s archival record, we quickly realized that the metadata schemas were not sufficiently expandable nor compartmentalizable. The archival record for the original manuscripts already contained metadata that pertained to its historical creation. When we started adding the names of transcription contributors, dates of the transcriptions’ creation, and an account of the transcription work with technical details, we inadvertently conflated the provenances.

Treating the transcriptions as data helped us address this provenance issue. While we ingested the born-digital derivative along with the scans of the original manuscript in our digital asset management system, we decided to also deposit and fully document the transcriptions in our Texas Data Repository. Within this preservation and long-term access system (DAMS), we were able to create a compartmentalized archival record with both provenances adequately reflected. Using the data repository also furthered our efforts to facilitate the reusability of this contributed knowledge in Digital Humanities work. Besides providing download access to a variety of text formats, the platform also mints a DOI, enabling the adequate citation of collaborative transcription work. To promote discoverability, we included this DOI under the “Related Resource” field of the original object’s archival record in our DAMS.

While FtP project coordinators expressed during the user study a need for various transcription export formats for future reuse and preservation, we found that basic non-proprietary file formats (i.e. TXT and CSV) were the most flexible and reusable. Although UT Libraries has been consolidating its digital assets into one DAMS, digital collections produced through external partnerships require separate platforms to maintain rights over digitized collection materials. As a result, our digital infrastructure consists of different content management platforms with distinct metadata schema and file format requirements. However, the identified non-proprietary formats are reusable across the board. From the perspective of digital humanities work, these formats are also human-readable and extensively used in popular visualization tools.

The symposium shed light on ethical approaches to reusing the results of crowdsourcing projects, especially in terms of crediting participants for their work and informing them about next steps. In terms of rights over the contributed transcriptions, we deemed the best approach would be to inform users in advance what rights statements would be applied to the results of the project. A panel about how to define a “meaningful” contribution to a project for the purposes of providing attribution raised a number of important points, including that for some volunteers anonymity is preferred. Projects favoring that approach used a blanket statement such as “transcribed by digital volunteers” where transcriptions appear in discovery platforms. Another important point raised was that there are many potentially valuable takeaways for project participants beyond having their name appear in a list of credits. Thoughtfully-constructed course assignments offer students mentorship and learning opportunities. Transcription events such as challenges or transcribe-a-thons can build community and generate citable, reusable materials. In all instances, we concluded that it is important to have clear communication with participants about the project’s purpose and what possible results are expected.

Final Thoughts

The technology internationalization work this project has accomplished has truly opened the door for non-Anglophone communities to engage in the Digital Humanities. Although the translation of the interface into Spanish and Portuguese might seem like a “drop in the bucket” of technology hurdles for these users, it has actually provided the foundation to address a similar obstacle for other linguistic communities. Among the ripple effects of this project is the current work in the development of a French interface dictionary that will easily plug into the revamped FtP infrastructure.

The attribution of collaborative transcriptions is slow but ethically imperative archival work. For far too long, researchers and cultural repositories engaging in the “crowdsourcing” of manuscript transcriptions have taken for granted the collective knowledge of citizen scholars. While we recognize that limited resources largely prohibit us from making transcription work compensated, the very least we can do is “give credit where credit is due” for arduous, sometimes interpretive, work that often becomes the foundation for other scholarship.

Bibliography

Bow, Catherine and Patricia Hepworth “Observing and Respecting Diverse Knowledge Traditions in a Digital Archive of Indigenous Language Materials.” *Journal of Copyright in Education and Librarianship*, vol. 3, no. 1, 2019, 1-36. https://doi.org/10.17161/jcel.v3i1.7485.

Caswell, Michelle, et al. “Diversifying the Digital Historical Record: Integrating Community Archives in National Strategies for Access to Digital Cultural Heritage.” *D-Lib Magazine*, vol. 23, no. 5/6, 2017, doi:10.1045/may2017-caswell.

Priani Saisó, Ernesto. “Las Humanidades Digitales En Español y Portugués. Un Estudio De Caso: DíaHD/DiaHD.” *Humanidades Digitales, Diálogo De Saberes y Prácticas Colaborativas En Red*, no. 12, 2015, pp. 5–18. *Tema Central*, Http://hdl.handle.net/10362/16145.

Risam, Roopika. "Decolonizing the digital humanities in theory and practice." In *The Routledge companion to media studies and digital humanities*, pp. 78-86. Routledge, 2018.