FromThePage Collection Owner User Study Report

Introduction

Enabling and Reusing Multilingual Citizen Contributions in the Archival Record is an NEH-funded project that focuses on supporting digital scholarship work in cultural institutions. Through a collaboration with Brumfield Labs, the project is expanding the features and access to FromThePage, a collaborative transcription, translation, annotation, and indexing platform. The work includes the implementation of additional export options that will facilitate the reuse of machine-readable textual outputs, and the creation of workflows to incorporate and credit citizen contributions in the digital archival record. To determine the specifics of this work, the project's team assessed the workflows and needs of the existing FromThePage (FtP) community through a user study of FtP collection owners in www.fromthepage.com and FtP instance managers. The survey was divided into three sections: import and export preferences, export access and preservation, and citizen collaborator attribution and rights.

Methodology

The primary investigator, Allyssa Guzman, the co-primary investigator, Albert A. Palacios, and the graduate research assistant, Joshua Ortiz Baco developed a Qualtrics survey with additional feedback from University of Texas Libraries Head of Assessment and Communication, Krystal Wyatt-Baxter, and the developers of FromThePage, Ben and Sara Brumfield. They created two blocks of questions based on users' response to whether they were individual researchers or cultural institution representatives. The former were asked about their use of the platform, indexing features, export preferences, and citizen collaborator rights. The cultural institution representatives were asked for additional information about the import features of FromThePage and preservation practices as related to their institutional priorities. Participants were identified by the FromThePage developers and contacted via email during April 2020.

Of the 27 respondents, 12 volunteers opted to participate in a follow-up 1-hour interview with the PI and co-PI to further discuss their practices regarding their projects' audience and practices for digital asset ingestion, FromThePage contribution reuse, and citizen collaborator attribution and rights. From the 12 volunteers, 7 participated in the interviews during June 2020. The PI and co-PI explained the purpose of the study to participants, asked about the four focus areas, and provided time for participants to ask questions. All participants agreed to have the audio of the interview recorded for the purpose of analyzing their responses once the user interviews were completed, at which point the recordings were destroyed.

Demographics

24 of the 27 survey respondents were library or archival staff that principally create and manage digital assets for their institutions in the FromThePage platform.[1] The remaining 3 participants' primary job function deals with public engagement and instruction.[2] The survey respondents were divided into two different groups (independent researchers and cultural institution representatives) based on their relationship with the institution holding the materials they ingested into their FromThePage project(s). For this survey, 23 out of the 27 survey respondents completely answered all the questions.

23 of the 27 survey respondents were library or archival staff, half of whom principally create and manage digital assets for their institutions. The remaining participants' primary job function entails public engagement and instruction.[3] Their stated principal use of FromThePage was to create transcriptions of manuscript collection materials from scratch or through OCR correction, with only 6 seeking to derive indexes from the materials and 3 hoping to create translations.[4]

Participant Categories

24 of these respondents created FromThePage collections using materials held and managed by their institutions. Their reported principal use of FromThePage was to create transcriptions of manuscript collection materials or perform OCR correction, with only 6 seeking to derive indexes from the materials and 3 hoping to create translations. Among the respondents, 3 reported using materials held and managed by institutions with which they have no affiliation.[5] All such participants indicated that transcribing materials was their primary use of the platform, but 2 are also interested in deriving indexes while 1 of them plans to create translations.[6]

[1] 12 of the participants identified their institution as a library, 6 as an archive, and 5 as a museum or another type of research institution.

[2] 5 of the participants identified as a director, a curator, a digital scholarship librarian, and a staff member with multiple roles, which the researchers have classified as being instructional or public engagement roles for this report.

[3] 5 of the participants identified as a director, a curator, a digital scholarship librarian, and a staff member with multiple roles, which the researchers have classified as being instructional or public engagement roles for this report.

[4] 24 of respondents identified the creation of transcriptions as their principal use of the platform.

[5] Within this group, one participant identified as an information professional who works at a library or archive and manages a FTP collection encompassing materials from several cultural institutions.

[6] Outlier participants of the study, instructors, were included in this group. They are outside the scope of this analysis since they tend to store research materials in their personal storage devices for subsequent editing and publishing, and so are less inclined to store their data in institutional repositories or consider them when exporting formats for their materials produced in FromThePage.

Results

Import and Export Preferences

22 collection owners imported their materials into FromThePage as a PDF/zip file upload, or from a IIIF or CONTENTdm repository. Interviewees pointed to the simplicity and accessibility of creating and uploading PDF and ZIP files locally without needing additional infrastructure. For those that did have digital asset management systems that could connect to FTP, they opted to use these import options, indicating the added benefit of being able to bring in some associated metadata.

Import Option	Number of Respondents[7]
Upload PDF or zip file	11
Import from CONTENTdm URL	6
Import from an IIIF repository	5
Import from the Internet Archive	2
Import from Omeka[8]	1
Create empty work	1

Currently, FromThePage collection owners can export transcriptions, translations, and indexes for individual works in Plain Text, TEI, HTML, IIIF/Open Annotation, and CSV formats, and for collection-wide indexes and transcriptions in CSV and XHTML formats. More than a quarter of the respondents indicated that they export plain text files to preserve alongside digitized assets.

Export Format	Number of Respondents[9]
Plain text file (.txt)	17

[7] Upon identifying the respondent’s affiliation to the institution holding the materials ingested into their FromThePage collection, the survey modified the number and type of questions asked to each participant. Since those not affiliated or employed by the institution holding their FromThePage collection materials represent an outlier group for this study, some questions, like this one, were not asked to them.

[8] Since this survey was created, FromThePage support for this functionality has been dropped in favor of using Omeka's IIIF support.

[9] This question was asked to the outlier group. All 3 expressed a preference for the Plain Text format, 2 found TEI-XML encoded file useful and at least one person mentioned CSV, both table and tags exports and XHTML as useful export formats.

IIIF/Open annotation	10
CSV table export	9
TEI-XML encoded file	9
XHTML	6
TEI encoded file	3
CSV tags export	3
None	2

When asked about desired additional export formats for item-level transcriptions, translations, associated comments, at least a quarter of the respondents expressed a strong preference for three specific outputs: a PDF containing the document image with a superimposed text layer (18), a PDF containing the document image and text side-by-side (17), and a Microsoft Word document (15). For indexes, collection owners expressed a desire for Microsoft Word, Microsoft Excel, or RDF formats. Interviewees considered these “low-tech”, familiar exports that could be easily used/reused by most users.

Currently, citizen contributors and visitors of FromThePage projects cannot batch download contributed content. When asked about the possibility of adding such a feature, collection owners were divided on whether or not they would enable user downloads of content. 12 of respondents were not sure about providing this access, while 9 felt comfortable enabling the features for their users.

Two main concerns emerged regarding the batch download of contributions by users: rights issues and the contributions’ quality. Foremost, collection owners were worried about providing access and enabling the commercial reuse/misuse of personal and sensitive information reflected in materials, specifically genealogical records. Others expressed concerns regarding “licensing implications” and the quality of the contributions. For one respondent, privacy issues could be addressed by the redaction of personal information prior to enabling users to download data. This need to review the contributed data connects to the second hesitation: the “cleanliness of data”, or the quality of the contributions. One collection owner and their colleagues wanted to maintain control over what items could be downloaded, permitting only the download of those that had undergone vetting and editing for accuracy.

Some collection owners saw the utility of the function, but only for specific users. For one, collection owners’ current ability to do so in the back end helps them assess project completion and conduct quality assurance. Another respondent considered the option to batch download “overkill for most users”, but a useful feature for “DH practitioners”. However, the general feeling is that FTP is primarily a collaboration management platform, not a repository from which users can download resources. Collection owners consider institutional repositories as the most appropriate method for servicing out contributions, ideally after they have been reviewed in some way. Considering this perception and that concerns are based on specific types of materials, perhaps the potential development of a batch download feature would be something collection owners would have to opt in, rather than opt out, as one respondent recommended.

Export Access & Preservation

In the survey, 14 respondents indicated that they ingested their FromThePage exports into their Digital Asset Management System (DAMS) as part of a digital object or record. They also deposited the exports into institutional repositories and used them to assist classes and provide access to researchers and the general public. These participants also declared that they reused FromThePage-created exports in digital humanities projects. Only 4 of all respondents mentioned using their exports as publishable materials. A small group was unsure of what the exported materials would be due to them being new users of FromThePage or because the project is not at an exporting stage yet.

Export Use	Number of Respondents[10]
Ingest into DAMS as part of the digital object/record	14
Provide access for community members	11
Provide access to classroom and researchers	8
Reuse in digital humanities projects	8
Deposit into institutional repositories	7
Serve as publishable materials	4
I do not export content created in FromThePage	0

FromThePage provides a non-default feature for collection owners to enable contributors to modify titles of individual pages in a document for use in metadata enhancement, but 15 of the surveyed indicated that they do not use this feature or that they are unsure about using it. Among the 3 that do enable title modifications, one individual incorporates this information within the metadata fields of “Identifier” and “Pagination”. For the metadata, the schemas used by most users are Dublin Core and MARC.

Metadata Schema	Number of Respondents[11]

[10] This question was not asked to the outlier group.
[11] This question was not asked to the outlier group.

Dublin Core	12
MARC	10
MODS	9
EAD	5
RDF	3
Unsure	2
Other:	1

In the case of indexes, 10 of the respondents indicated that they incorporate indexes as metadata of the digital asset or archival record; 6 store them as datasets that are used to create visualizations of the collection or object; and 3 treat them as stand-alone digital objects. Two participants do not use the indexing feature in their projects.

16 participants described the accuracy of contributions as important for their projects; 15 stated that they regularly check them for quality and correctness; one respondent mentioned that accuracy was moderately important; and 2 checked contributions occasionally.

Citizen Collaborator Attribution & Rights

15 of the surveyed indicated that they credit citizen contributors. Of these, 9 of them use their collaborators' FromThePage username for attribution while 6 use full names. Other participants shared alternative methods of attribution such as crediting all collaborators as a unified group and asking citizen collaborators to sign a General Data Protection Regulation (GDPR) informed consent declaration. One respondent expressed doubt as to how to credit citizen collaborators.

The most popular methods used to include attribution were: including contributor information within the metadata records, within the FromThePage collection's description, and displaying the attribution in the project's platform. An equal amount of 4 participants indicated preference for each of the previous attribution methods. One person gave credit within the contribution itself. On the subject of Rights, only one participant reported having a contribution conditions or licensing policy for citizen contributions. The majority of participants (9) do not have this policy and 2 were not sure.

Among the participants who do not offer any credit attribution to citizen collaborators, 5 offered several reasons to not do so. One participant explained that their FromThePage project is new and so they have not thought about this issue yet. Another 2 clarified that their projects are not open for citizen collaborators; one of these indicated that only staff members work on it. One person opted to clarify that the transcriptions were crowdsourced instead of giving individual attribution, and one last participant said that collaborators have not asked about attribution and so the participant implied that citizen collaborators are not concerned about it.

Findings or Discussion

Import & Export Preferences

- PDF and zip files are the reported preferred method of ingestion due to their simplicity. They were followed by CONTENTdm and IIIF manifests.
- For export formats, plain text was found to be the most useful export format.

Export Access & Preservation

- The most popular preservation method is ingesting the exports created in FromThePage into digital asset management systems (DAMS) while providing access to the community in some way that was not specified.
- Dublin Core and MARC were considered the most popular metadata schemas used, and asset accuracy was generally important to survey participants.

Citizen Collaborator Attribution & Rights

- More than half of the respondents credit their citizen contributors, but there was not one single preferred method for doing so among the several methods mentioned.
- For rights and usage policies, Creative Common Licenses were identified as the most common and easiest way of giving credit and providing access to contributions. However, most participants have not yet included such licenses or policies in their projects.

Future Steps or Recommendations

Participants in the user study were generally in agreement in their preferences and recommendations about proposed improvements to the FromThePage platform as outlined above. All the participants also expressed satisfaction with the current features of the platform and willingness to test and incorporate new features. Stemming from their input, we recommend the following as potential future developments of the platform:

User-facing elements and additional settings

- Because the images that are uploaded to the platform can be of any quality, give users the ability to adjust the image settings (saturation and contrast, for example) to help them improve the visibility of the texts or details in the images.
- Provide users the ability to edit and contribute to the metadata fields and simplify the process by allowing users to copy and paste information from the transcription or translation into the corresponding metadata fields.

- Provide citizen contributors the option to add their ORCID to their profile so that it can be added to the contribution as metadata.[12]
- Enable simultaneous collaboration to enhance classroom and workshop use of the platform.

Owner Dashboard

- Include an auto-populate feature in the item's metadata that shows the usernames of citizen contributors that worked on the transcription of a work. Ideally, this feature would also show a "completion" percentage for each citizen contributor to help collection owners determine who to credit for "substantial/significant" contributions.
- Most owners expressed interest in adding Creative Commons Licenses into their projects, but did not specify which ones were best suited to their needs. Hence, we recommend that a drop-down menu is incorporated into the "Settings" tab of the collection and the "Start a Project" tab that allows owners to select whichever is most appropriate to their project.
- Collection owners were also interested in being able to manage licenses by project. Enable a pop-up window that forces citizen contributors to agree to the license stipulations for a given project at the moment of joining a project.
- Another enhancement suggested was the inclusion of an option through which the transcription instructions could be applied globally or individually depending on the collection owner's preference.

Export features

- Participants highlighted the importance of familiar and simple-to-use export formats. like plain text, PDF, and DOC files. Additional exports, like HTML and markdown files with limited formatting and links (IIIF manifests or deep links), would also be important for reuse in digital publishing and static-site projects. There is also a need to include variations to the current exports, including PDF with transcription or translation text side-by-side on the same page in landscape orientation; PDF with a layer overlaid of text; and plain text files with transcription and translation
- It would be ideal if exporting could be done at the page, item, and collection level. Additionally, we also recommend a feature for owners to batch export a selection of works within a collection and an option for selecting a range of pages within individual works.
- Collection owners would also benefit from the creation of an option to enable users to export their own work, transcription or translation, but limiting the exporting of other users' work.

[12] Since this report was created, Brumfield Labs has added the ability for users to add their ORCID to their profile.

- The “Export” would include an additional option to include metadata within the pages of the export. Collection owners would then select if metadata fields would appear within the pages and which ones.
- Participants also expressed an interest in having the option to modify file names and metadata prior to export to facilitate post-processing of files. This would be an additional feature in the “Export” tab in the collection page that would work as a staging area prior to exporting and would include options to edit metadata at the page, item, and collection level. Each of these levels would have shared metadata fields, such as creator or date, and independent fields pertaining to each page within an item, and to each item within a collection.
- Add an export feature to automatically generate a list of the usernames of people that contributed to the particular pages or collections, similar to the current “Mailing List Export” feature.