TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

INTERIM REPORT
FOUNDATION PROJECTS

ENGAGING CROWDS
Citizen research and cultural heritage data at scale

PI: Pip Willcox, The National Archives, UK

The National Archives, UK | Royal Botanic Garden Edinburgh | Royal Museums Greenwich | Zooniverse, University of Oxford

DECEMBER 2020

TABLE OF CONTENTS

Executive Summary

Project summary

The last two decades have seen increasing growth in digitally enabled volunteer participation in cultural heritage. *Engaging Crowds: citizen research and heritage data at scale* Towards a National Collection (TaNC) Foundation Project explores its current and potential practice, through three strands of activity. The project is led by The National Archives, with Royal Museums Greenwich, Royal Botanic Garden Edinburgh, and the Zooniverse team at the University of Oxford.

Research methods

We are building an indexing tool for the Zooniverse Project Builder which enables volunteers to steer their own path through a project. We are delivering three citizen research projects to surface the contents of record sets held by the project's Independent Research Organisations and to evaluate the tool: *HMS NHS: The Nautical Health Service*, *Scarlets and Blues*, and *RBGE Herbarium: plants, collectors and discovery*. We are surveying the use of citizen research in cultural heritage, and exploring through workshops the re-use of data produced through citizen research projects by volunteers, by collections-holding organisations, and by Artificial Intelligence, and will share our findings with the wider community.

Progress

Our work to date has focused on the design of the indexing tool and of the three citizen research projects which will launch early in 2021. We have run two project workshops with international participants, including from around the UK, on the use of automation (AI, such as machine learning) at all stages of citizen research projects, and on the perspective of cultural heritage organisations.

Covid-19 impact

Our immediate project team has been affected by the pandemic including through furloughing, but we have moved work packages within the project to avoid delays to the overall scheduling. We have made accommodations such as taking a workshop online, and have used the opportunity to increase international participation. Given the wider effects of the pandemic on the sector, we may have limited engagement from cultural heritage organisations who lack capacity to share their findings and experiences of running citizen research projects. In this case, we will focus our research on the published record.

Next steps

Work on the indexing tool continues, refining it according to the needs of each citizen research project.

These projects will launch in sequence in early 2021 and their volunteers will be supported by project members. In due course, data from these projects will be analysed to evaluate the success of the indexing tool in increasing volunteer engagement.

Work is underway to survey the use of citizen research in cultural heritage organisations, and to survey the volunteers themselves.

We are writing up our first two workshops, and will deliver the final workshop in autumn 2021.

The platform for sharing data from the projects will be designed and built by Zooniverse for winter 2021, drawing on advice from volunteers.

We will share our findings with the community through presenting the project at conferences, regular blog posts, submitting a peer-reviewed article, and delivering the project report on cultural heritage citizen research.

Abstract

The *Engaging Crowds: citizen research and heritage data at scale* TaNC Foundation Project explores the current and potential practice of engaging diverse audiences with cultural heritage collections through the creation, use and reuse of heritage data. The last two decades have seen a revolution in volunteering programmes, as cultural heritage organisations have adopted digitally enabled approaches to crowdsourcing, and this project is part of that wider landscape. The project is led by The National Archives (TNA), with the Zooniverse team at the University of Oxford, Royal Botanic Garden Edinburgh (RBGE), and Royal Museums Greenwich (RMG).

Our project has three focuses: community consultation on citizen research in cultural heritage organisations, including through workshops; prototype tool development; and evaluating the tool through three citizen research projects and analysis of their data. The project engages with the wider community through seeking volunteers for the three citizen research projects and working with them once the projects launch; through our three workshops; through conferences and workshops; and through asking for information about previous cultural heritage projects that used digitally enabled citizen participation. A report based on these findings, a review of current practice, and the outcomes of the series of workshops will recommend best practice in encouraging and supporting meaningful public interaction with heritage collections. This work informs the mission of TaNC programme, enhancing our understanding of engaging the public digitally with our cultural heritage.

Aims and Objectives

This collaborative project will help us take steps towards a unified national collection by identifying and addressing current and future challenges facing the effective conduct, use, connection, and re-use of citizen research and the data it produces. It has the following objectives.

1. Understanding the current state of citizen research in heritage organisations

Volunteer participatory research has been used to gather data to enrich and inform scientific enquiry for at least two centuries. In our digitally and internet-enabled world, our ability to share the subjects of our research and to increase the pool of volunteers who can lend their time and analytical skills to projects have transformed citizen research. This brings the potential to enhance meaningful access to collections and draw on the skills and knowledge of a range of participants as diverse as our society. Through a review of published research, blog-posts and unpublished reports on cultural heritage citizen research projects, and through studying case-study projects we will seek and analyse evidence on who participates in this work, what their motivations are, and work towards understanding how heritage organisations can enhance that experience to increase participants' enjoyment and levels of engagement.

2. Create a prototype indexing tool to enable navigation of research subjects in citizen research projects

When analysing a series of images, the order in which they are served to volunteers is of low importance; when volunteers are tasked to engage with textual content, the ability to navigate their own paths through content has been identified as key to maintaining interest in participation. Our project is developing an indexing tool that will enable this navigation. It will be evaluated through the three case-study projects and a workshop dedicated to the volunteer experience. The indexing tool will trial this self-navigation method as one means of engaging participants more deeply in individual projects.

3. Explore barriers to and identify solutions for the effective use and re-use of citizen-research produced data

Increased data production brings maximum benefits when it is productively used by all its potential audiences. We have identified three audiences for the data: collections-holding organisations,

research communities including Artificial Intelligence (AI) and machine learning, and the public including researchers and industry.

Can collections-holding organisations use the data to enrich their understanding of the collections, including new knowledge in cataloguing, interpreting and linking collections while maintaining public trust in the reliability of the tools they provide?

Can AI and machine learning research communities collaborate to increase the automation in citizen research projects, ensuring participants' time is increasingly used on tasks that require human skills while the machines learn from them?

Can the public access the data volunteers produce, using appropriate tools and skills to interrogate, link, interpret and repurpose that data?

Workshops including representatives of these audience groups will inform this exploration.

4. Collection, analysis, and dissemination of evidence on citizen research, with recommendations to inform policy for the development of a future national and supranational citizen research effort

Having gathered data on the current use of citizen research by cultural heritage organisations, on barriers and solutions to the re-use of data produced by volunteers, and tested the potential to increase meaningful engagement with projects through the indexing tool, we will produce summary and in-depth reports to share the project's findings with the wider heritage and policy communities, including recommendations for future methods and developments. Working with the TaNC programme and its other Foundation Projects, we will feed into a co-created vision and roadmap for a connected virtual national collection.

Partnership structure

The National Archives

TNA is leading the project, its reports and surveys, and dissemination of project outputs. It is creating, supplying image data and metadata, providing records expertise, and running a citizen research project on the Royal Hospital Chelsea’s records, and organising a workshop on automation.

The principal investigator, project manager, and communications officer are based at TNA. A records specialist and citizen research project designer are delivering TNA’s citizen research project, with input from a research associate, a research fellow in citizen research at TNA, and a project design advisor who had a student placement with TNA during the design phase of the project.

Royal Botanic Garden Edinburgh

RBGE is creating, supplying image data and metadata, providing records expertise, and running a citizen research project on the plant family Gesneriaceae, and organising a workshop about collections-holding organisations. Colleagues are contributing to project reports and surveys, and to dissemination of project outputs. A co-investigator and project officer are based at RBGE.

Royal Museums Greenwich

RMG is creating, supplying image data and metadata, providing records expertise, and running a citizen research project on the Dreadnought Seamen’s Hospital records, and organising a workshop on the volunteer experience and public re-use of data. Colleagues are contributing to project reports and surveys, and to dissemination of project outputs. A co-investigator, research consultant, and super volunteer are based at RMG.

Zooniverse

Zooniverse is building, implementing and iterating the indexing tool for citizen research projects and a data sharing platform for citizen research projects. Colleagues are contributing to dissemination of project outputs. The development manager, projects consultant and developer team are based at Zooniverse.

Staffing Structure

Pip Willcox, Head of Research, TNA — Principal investigator. Project direction, leads on reports and surveys, on TNA's citizen research project delivery and workshop on automation, ensures delivery of project to proposed timeline and budget, liaising with TaNC Programme Director and other Foundation project PIs.

Chris Lintott, Professor of Astrophysics, Co-founder of Zooniverse, University of Oxford — Co-investigator. Leads on indexing tool and data platform development, contributes to outputs and dissemination.

Elspeth Haston, Deputy Herbarium Curator, RBGE — Co-investigator. Leads on RBGE's citizen research project and workshop on collections-holding organisations, contributes to outputs and dissemination.

Martin Salmon, Research Curator & Archivist, RMG — Co- investigator. Leads on RMG's citizen research project design and delivery, and on volunteer experience workshop, contributes to outputs and dissemination.

Sam Blickhan, Humanities Lead, Zooniverse — Development manager. Manages the development and implementation of the indexing tool and data platform, advises on citizen research project design and workflows, and community engagement.

Grant Miller, Communications Lead and Community Manager, Zooniverse — Projects consultant. Advises on citizen research projects including design and workflows, and community engagement.

Bernard Ogden, Research Software Engineer, TNA — Project designer. TNA citizen research project design, data workflow engineer for ingestion into TNA's catalogue, contributes to volunteer engagement, contributes to project website content and maintenance.

Will Butler, Head of Military Records, TNA — Records specialist. Leads on records, providing expertise to the TNA project team, contributes to volunteer engagement.

Andrea Kocsis, Friends of The National Archives Research Fellow (Advanced Digital Methods), TNA — Research associate. Contributes to research into volunteer engagement with findings from her one-year research fellowship (November 2020 – October 2021).

Thomasina Smith, Placement Student, TNA — Project design advisor. During a four-week placement (August 2020), collaborated on TNA citizen research project workflow design.

Sally King, Digitisation Officer/Herbarium Volunteer Coordinator, RBGE — Project officer. Implements RBGE citizen research project, contributes to volunteer engagement, implements workshop on collections-holding organisations.

Stuart Bligh, Head of Research and Information, RMG — Research consultant. Advises on RMG citizen research project and liaison across RMG.

Trevor Nash, Volunteer, RMG — Super volunteer. Advises on RMG citizen research project workflow design and volunteer experience.

Louise Seaward, Academic Engagement Manager, TNA — Project manager. Manages all administrative aspects of the project, advises on workshop design.

Liz Fulton, Academic Communications and Impact Officer, TNA — Communications officer. Manages project communications, including project website and liaison with TaNC.

Covid-19 Impacts

While Covid-19 has meant we moved activities within the project, the overall schedule remains unchanged, with no extension requested.

All project partners have had to deal with the challenge of moving to full-time remote working. This has been particularly challenging for work which involved coordinating between developers and project teams, adding additional complexity to the tasks of beginning the build of the tools. Members of the project team have been furloughed, or had working hours reduced for certain periods.

All project meetings have taken place online, as the first in-person meeting was scheduled for a date after the announcement of the first lockdown.

The first workshop took place in person, in December 2019. The second workshop was taken online in December 2020. To avoid 'Zoom fatigue' we decided to reduce its duration to 2.5 hours. RGBE and TNA put significant work into designing the online workshop to make the most of its format which allows for higher attendance and participation from further afield. These aspects of the workshop will be shared, and will feed into the planning of the third workshop which is scheduled for September 2021, and is likely to be held online.

Response to our requests for information about cultural heritage citizen research projects has been limited to date. We hoped colleagues would be willing to share internal reports or evaluations about their experience. We have allowed a longer period for this stage of data gathering. Given the pressure this year has put on colleagues and organisations in the sector, it is possible we will not receive many responses to our call. In this case, we will work with the information we have, and focus our efforts on published reports, blog posts, academic publications etc.

Reports suggest that there has been increased activity in citizen research projects online during these times of restrictions to movement and in-person socialising. It could be that a side-effect of the pandemic is more volunteers choosing to donate their time to our citizen research projects once they launch.

Revised Overall Programme

Date	Milestones	Work package
Dec 2019	TNA hosts workshop on machine learning and citizen research - prior to project launch	WP2
Feb 2020	Project start; set up project management tools Zooniverse begins work on indexing tool & data platform	WP1 WP3, WP4
Apr 2020	RMG, TNA & RBGE begin citizen research projects design	WP5
May 2020	Engaging Crowds website goes live	WP6
Oct 2020	Zooniverse delivers prototype indexing tool Testing phase begins for RMG citizen research project Work on narrative report starts	WP3 WP5 WP2

	Project advisory board meets	WP1
Nov 2020	Testing phase begins for TNA citizen research project Zooniverse tests and refines indexing tool	WP5 WP3
Dec 2020	RBGE hosts workshop on data management and re-use	WP2
Jan 2021	RMG citizen research project launches Start survey of volunteers Testing phase begins for RBGE citizen research project Zooniverse refines indexing tool	WP5 WP4 WP4 WP3
Feb 2021	TNA citizen research project launches Present paper at conference	WP5 WP6
Mar 2021	RBGE citizen research project launches	WP5
Sept 2021	RMG delivers final workshop on volunteer experience Delivery of report on citizen research landscape Project advisory board meets	WP2 WP2 WP1
Oct 2021	Survey of volunteers across all three projects closes Citizen research projects close RMG, TNA, RGBE start data aggregation and quality assurance	WP2 WP5 WP5
Nov 2021	RMG, TNA and RBGE analyse survey and citizen research data Zooniverse delivers data sharing platform	WP2, WP5 WP4
Jan 2022	Report delivered and project ends	WP6

Events and Consultations

Date	Event	Subject	Attendees	Notes
13 Dec 2019	***People and Machines - co-creating with heritage collections.*** Workshop hosted by The National Archives	How to combine machine learning and citizen research in cultural heritage	44	With attendees from the US, across Europe and the UK, the workshop discussed the current use, potential, ethics and practical blockers of using AI such as machine learning at all stages of the citizen research workflow. TNA is synthesising and analysing the outputs of the six discussion groups and will publish the report.
Jun 2020	Engaging with cultural heritage organisations and groups in a call for information	Requesting information on experiences of citizen research	TBC	This work is ongoing.
1 Dec 2020	***After the crowds disperse: crowdsourced data rediscovered and researched*** Workshop hosted by Royal Botanic Garden Edinburgh (online)	Flow of data from citizen research projects back to collections-holding organisations, including quality control, ingestion and data reuse	60	The workshops discussed ideas on best practices for citizen research projects to promote the existence of and access to collections, methods of quality control and analysis of the resulting data to ensure that the results can be used (and re-used) effectively. RBGE will lead on the report based on notes created by participants during the event.
Sept 2021	Final workshop hosted by Royal Museums Greenwich	For community researchers, testing & using Zooniverse Project Builder & indexing tool	TBC	A report on workshop findings will be shared.

Research Approach

Public participation in heritage research has the potential to engage new audiences, to enlist volunteers in analysing and generating data at scale, and to invite new perspectives on our national collections. Key to releasing this potential is effective engagement of diverse audiences, and the development of workflows for the creation and re-use of data within collection discovery platforms, for training automated systems, and to give access to the citizens and researchers.

We will explore ways of extending and deepening engagement across communities, proposing a best-practice framework for future citizen research projects with heritage data, informing their design and modelling.

The project is drawing on expertise from across sectors to expand our citizen research community and to ensure the effective re-use of crowdsourced data by addressing the following questions.

1. How can we best engage volunteers across the nation's communities with citizen research projects, to further a shared understanding of our collections? What existing methods and data are the most successful for measuring that engagement?
2. How does the ability to navigate one's own path through the data of a citizen research project affect engagement with the project?
3. How can we verify, assess, present, and value the contributions of citizen research?
4. How can we enable the re-use of crowd-sourced data within collection discovery platforms, for training automated systems, and to give access to citizens and researchers that supports and encourages further engagement, re-use and analysis?
5. Does easy access to data created by citizen research projects affect engagement with projects? What other tools are necessary to enable meaningful access to this data?

We are interrogating these questions within the project as we build and refine three citizen research projects on the Zooniverse platform and run by RMG, TNA and RBGE, with the integration of a new indexing tool created by Zooniverse developers. We are also reaching out to our audiences in and beyond the heritage sector, including those within other TaNC Foundation Projects, asking them to share written reports of their citizen research experiences, and discussing challenges and solutions in a series of workshops.

Project tasks are organised under six work packages:

- WP1 Project management
- WP2 Research and discovery
- WP3 Indexing tool prototype development
- WP4 Platform development
- WP5 Citizen research projects
- WP6 Dissemination

Early Research Results/Outputs

To date the project has concentrated on tool development and designing three citizen research projects. The Zooniverse team have been designing, building and iterating a new indexing tool that will give citizen researchers agency to choose their own path through these projects. The projects are run by RMG, TNA and RBGE on the Zooniverse platform.

Indexing tool prototype development (WP3)

The Zooniverse team have planned development of the indexing tool as a three stage process, to coincide with the building and launch of the three individual citizen research projects from RMG, TNA, and RBGE respectively.

The Zooniverse team have been supporting each partner as they create their new citizen research project using the Project Builder tool. For each project, the Zooniverse team has organised multiple sessions to demonstrate how the platform works, help test workflow prototypes, assist with data uploads, and generally support the research teams with all aspects of creating their Zooniverse projects. The expertise of the Zooniverse team has been invaluable in this.

The Zooniverse designer has created a clickable InVision prototype as well as user stories to help guide the active development process for the indexing tool. Zooniverse collected example

images/metadata from each team to help get a sense of what type of information volunteers will be able to search/sort on (e.g. date, author, and location). Additionally, Zooniverse included a short focus on developing the dropdown menu feature, as it was quickly identified as a tool that all three projects would benefit from using.

Attention is currently focused on RMG's project, where the indexing tool will be deployed for the first time. For the first citizen research project, the major aim of the indexing tool is to deliver specific sets of 'subjects' (images of text) to volunteers, and allow them to choose which sets to work on. Within each set, volunteers will be able to work sequentially through the 'pages' within that set of uploaded images. Once the RMG project is fully set up, the Zooniverse team will continue to iterate on the indexing tool, to meet the stated development goals as well as incorporate user feedback, to be implemented for the projects at TNA and RBGE.

Citizen Research Projects (WP5)

Each of the three projects aims to uncover the content of records held by the cultural heritage partners. The projects will launch in sequence from early 2021 and appear here in launch order.

Royal Museums Greenwich

The RMG citizen research project, *HMS NHS: The Nautical Health Service*, is based on the records of Dreadnought Seamen's Hospital (RMG, DSH), a hospital for merchant seamen that existed in Greenwich for over 150 years (from 1826–1986), and the main clinical site for seafarers entering or leaving the busy port of London from all over the world. The records are kept at the Caird Library of the National Maritime Museum, Royal Museums Greenwich, and have been digitised by the genealogy company Ancestry. Ancestry's interest was in indexing family history details from the records only. Making the entirety of the records available digitally would offer researchers the chance to analyse their medical information, to see what trends and patterns are evident in over 100 years' worth of data through a very large number of nineteenth- and early twentieth-century case studies in the history of medicine and seafaring.

RMG has created the project using the Zooniverse Project Builder, trialling different workflows and developing instructional guidance. The format of the original Admission registers is 12 columns across a double page spread. Each row across a double-page contains information about one patient, with each column featuring the same type of information about different patients. Early iterations of the workflows experimented with transcription by columns (quicker and the preferred choice of RMG's 'super volunteer') or by rows (potentially more interesting for volunteer transcribers). As both were time consuming, a third option was trialled: a separate workflow for each individual column, breaking a large task down into manageable chunks. Reflecting all three approaches, a number of separate workflows were devised and tested by the *Engaging Crowds* team and a select group of existing volunteers at RMG during the project's alpha phase. RMG worked with Zooniverse to incorporate feedback received from the alpha testing into an improved version of the project. The beta phase of the project will be initiated in January 2021, where Zooniverse volunteers will be invited to test the revised project and share their feedback.

The National Archives

TNA's forthcoming citizen research project, *Scarlets and Blues*, takes a behind-the-scenes look at the lives of people at the Royal Hospital Chelsea during the First World War (TNA, WO 250). The project title references the two types of uniform coat worn by the Chelsea Pensioners who live in the Hospital.

TNA has developed the design for *Scarlets and Blues* using the Zooniverse Project Builder. The records occur in two types: the minutes of meetings, and an index into those meetings with information presented in columns. The index also makes significant use of annotations: arrows, punctuation and other marks used to connect pieces of text. This necessitated the building of two

workflows with accompanying guidance documentation, for each type of page contents, with the *Index* workflow orientated towards collecting information, and the *Minutes* workflow designed for transcription.

TNA has also considered whether Handwritten Text Recognition (HTR) could be a part of the workflow, effectively replacing the 'transcribe' stage of the *Minutes* workflow with a 'correct' phase. Results of automatic transcription with a Transkribus model trained on a very few pages were better than expected. This suggests that such an approach might be feasible, but as there is not a clear way to correct text within Zooniverse workflows, this approach has been set aside. If time permits, we will use the manual transcriptions to experiment with training a better HTR model that could be used for other minute books of the Royal Hospital Chelsea, or other digitised records from the period written in similar hands. We will share these training models.

In November 2020 TNA completed alpha testing of the project, receiving feedback from members of the *Engaging Crowds* project team. TNA is now responding to the feedback and readying the project for its beta testing phase, where it will be opened to participation and comments from the Zooniverse community.

Royal Botanic Garden Edinburgh

The RBGE citizen research project aims to transcribe the data on labels of plant specimens collected from across the world over a period of more than 200 years. The specimens, with their collection labels can be used to answer many research questions including on changes in species distribution and flowering times linked to habitat loss and climate change. The specimens in this project are all in the Gesneriaceae plant family, familiar to many from the widely cultivated houseplants, African Violets.

RBGE initially focused on data structure and biodiversity data standards. The team has identified the data to be transcribed, based on the Minimal Information about a Digital Specimen (MIDS) standard. RBGE is one of the institutes leading the development of this standard within the international Biodiversity Information Standards community. RBGE aims to capture data to MIDS level 2 which provides the data required for most research purposes.

RBGE's Zooniverse project, The *RBGE Herbarium: plants, collectors and discovery*, exists with a test dataset of specimen images. RBGE is currently building the workflows, tasks and guidance documentation, including tutorials. Zooniverse has provided support by giving access to additional functionality, including dropdown lists and a 'combo' functionality to group tasks on a single page. RBGE has also started to develop linked hierarchical dropdown lists within the locality data combo task. The plan is to create two projects using the Zooniverse Project Builder: a baseline project and indexed project, splitting up our digital specimen images of the plant family Gesneriaceae into two randomised datasets. These will build upon previous analysis of data quality from citizen research projects carried out at RBGE. We aim to analyse how this impacts the both transcriber engagement and data quality.

Dissemination (WP6)

Our communications officer, Liz Fulton, has designed an overarching communications strategy for the project, involving a combination of social media, blog posts and volunteer outreach. Joe Padfield (The National Gallery) generously set up GitHub-based websites for the TaNC Foundation Projects and we have populated our website with project information.

Each partner will deliver a blog post on project updates, to appear regularly across the life of the project. TNA and RMG have produced the first blog posts. We decided against starting a social media account for the project that would fall into disuse after its two-year span and have instead been using our institutional social media accounts to promote project news. *Engaging Crowds* was presented at two TaNC Discovery Webinars in August 2020.

Draft schedule for project blog posts

Date	Partner responsible	Suggested topic
July 2020	TNA	Introduction to project
Sept 2020	NMM	Introduction to citizen research project
Jan 2021	RBGE	Report on workshop
Jan 2021	RMG	Launch of citizen research project
Feb 2021	TNA	Launch of citizen research project
Feb 2021	Zooniverse	Overview of indexing tool
Mar 2021	RBGE	Launch of citizen research project
Oct 2021	RMG	Report on workshop
Feb 2022	TNA	Project summary

Advisory board

Our Advisory Board will meet twice during the project to guide our work and dissemination. We are very grateful to the Board members for their generosity in sharing their time and their expertise. They bring valuable knowledge and experience from different perspectives around citizen research.

Adam Corsini, Collections Engagement Manager, Jewish Museum London

Stuart Dunn, Reader in Digital Humanities, Department of Digital Humanities, King's College London

Libby Ellwood, Global Communications Manager, iDigBio

Siobhan Leachman, Citizen Scientist and Wikimedian

Next steps

Research and Discovery (WP2)

TNA is conducting a literature review and liaising with Independent Research Organisations and other cultural heritage organisations and groups in the UK and beyond to gather publications and data relating to cultural heritage citizen research projects for a report on the state of citizen research in the sector.

We will survey the volunteers who take part in each project to understand more about their motivations, preferences and engagement.

Our final citizen research workshop is planned for September 2021, to be hosted by RMG. It will focus on testing the Zooniverse indexing tool and seeking feedback from volunteers and collection managers. It will canvas the views of volunteers on their motivations, experiences, and their re-use of data produced through citizen research.

Indexing tool and citizen research projects (WP3 and WP5)

RMG will launch its citizen research project and engage with its volunteer community. TNA will integrate volunteer feedback into their project before launching it. RBGE will test their baseline project, and launch it in March 2021; the indexed version of the project will follow. Zooniverse will continue to support partners with the design and delivery of their projects and develop the indexing tool to meet the needs of each project.

Best practice suggests that images used during citizen research projects as well as the classification or transcription data should be made available. This is not always permissible under cultural heritage organisations' licensing terms. We continue to investigate the possibilities in this area.

Platform development (WP4)

A platform for sharing the data created by citizen research projects will be designed and implemented, facilitating access to these projects' outputs as an additional way to engage volunteers and credit their work publicly. This will be hosted on the Zooniverse platform.

Dissemination (WP6)

We are planning to present our work on *Engaging Crowds* at several conferences. We will be delivering a paper at the 5th AHRC Connected Communities Heritage Network Symposium in February 2021. We will also be submitting an abstract to deliver an interactive session at the DCDC conference, which will take place in June 2021.

We hope to present papers at digital humanities and cultural heritage conferences. If these conferences do not go ahead due to the pandemic, we will look for alternative venues.

Our work will be shaped into a final report and an article, which we plan to submit to a peer-reviewed journal at the end of the project.

Contacts

Louise Seaward, *Engaging Crowds* Project Manager, The National Archives

louise.seaward@nationalarchives.gov.uk

Pip Willcox, *Engaging Crowds* Principal Investigator, The National Archives

pip.willcox@nationalarchives.gov.uk

Annexes and links

Project website

https://tanc-ahrc.github.io/EngagingCrowds/

Zooniverse repository

https://github.com/zooniverse

Scarlets & Blues

Aggregation code (still in development): https://github.com/bogden1/s-b_aggregation

Handwritten Text Recognition (HTR) Experiments

We carried out small, time-bound experiments using Transkribus, building train and test sets from 16 transcribed pages. This is lower than the number of pages normally recommended for training a Transkribus model. The HTR + 'English Writing M2' model was used as a base and training was over 50 epochs.

In the first experiment, 13 pages were used to train, achieving a character error rate (CER) of 13.43% on the validation set (of just 3 pages).

In the second experiment, 15 pages were used to train, achieving a significantly improved CER of 4.95% on the validation set (just 1 page, with quotation marks now removed from the transcript of that page).

These CERs are not meaningful, as the validation sets are far too small. A brief and unscientific 'eye test' suggests that transcribed pages *may* be good enough quality for a manual correction to be more effective than full transcription. This encourages us to experiment with training a model with a larger train and test set of transcriptions. As noted above, for the work of this project we are not pursuing HTR as a method of producing transcriptions. Rather, a potential re-use of the data transcribed by volunteers in *Scarlets and Blues* is as training data for future work.

Transkribus: https://readcoop.eu/transkribus/