TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

FINAL REPORT

FOUNDATION PROJECTS

Engaging Crowds: citizen research and cultural heritage data at scale

PI: Pip Willcox, The National Archives

The National Archives | Adler Planetarium
Royal Museums Greenwich | University of Oxford
Royal Botanic Gardens Edinburgh | Zooniverse

October 2022

TABLE OF CONTENTS

Authors

Louise Seaward, Pip Willcox (The National Archives), Samantha Blickhan (Zooniverse-Adler Planetarium), Stuart Bligh (Royal Museums Greenwich), Will Butler (The National Archives), Liz Fulton (The National Archives), Elspeth Haston (Royal Botanic Garden Edinburgh), Ashleigh Hawkins (The National Archives), Rebecca Hutcheon (The National Archives), Sally King (Royal Botanic Garden Edinburgh), Andrea Kocsis (The National Archives), Chris Lintott (Zooniverse-University of Oxford), Grant Miller (Zooniverse-University of Oxford), Trevor Nash (Royal Museums Greenwich Volunteer), Jim O'Donnell (Zooniverse-University of Oxford), Bernard Ogden (The National Archives), Martin Salmon (Royal Museums Greenwich), Thomasina Smith (The National Archives Undergraduate Student Placement)

Executive summary

The *Engaging Crowds* project explores citizen research[1] in cultural heritage: people using digital, usually web-based technologies to contribute to knowledge about collections. At its best, citizen research can create a virtuous circle of participation and shared new knowledge about our national collection. Members of the public can engage with collections deeply and richly, developing and sharing their expertise and building meaningful relationships with our shared heritage. Cultural heritage practitioners can select collections of focus, including those that have been underrepresented or overlooked. Heritage organisations benefit from the new insights produced through these projects, feeding new knowledge into shared digital resources such as catalogues or digital archives.

This virtuous circle requires energy to keep its momentum. Crowdsourcing relies on the generosity of citizen researchers donating their time and skills to projects. It relies on collections and digital infrastructures, including people's expertise, for crowdsourcing platforms, collections digitisation, project design, ongoing engagement with the public throughout the life of a project, and undertaking to use the data that is produced well.

We engage in crowdsourcing because we care about the collections and the community. This work, paid or voluntary, involves commitment and emotional labour. The ethics of crowdsourcing must be at the front of our minds: ethics inform our values which are evident in every aspect of our work, from first reaching out to a community to sharing the final data. Citizen research is not a 'cheap option for cataloguing'; it is an opportunity to develop and nurture sustained relationships of mutual benefit between collections and the public and requires ongoing resources from across organisations.

To create "a unified virtual 'national collection'"[2] we need to create and understand collections data so we can make links and create new knowledge. Other Towards a National Collection (TaNC) Foundation and Discovery Projects explore aspects of automation; *Engaging Crowds* explores how citizen research can help with this. By definition, crowdsourcing supports TaNC's aim to increase "public access beyond the physical boundaries" of collections locations, and can contribute towards becoming more inclusive, redressing the demographic and geographic imbalance in cultural heritage.[3]

Project summary

Our project investigated citizen research in the round, seeking perspectives of many stakeholders through a range of methods. We surveyed the current landscape of cultural heritage crowdsourcing, particularly across the UK. We developed an indexing tool to enable citizen researchers to choose their own pathway through a project on the Zooniverse platform (project partners, and the largest crowdsourcing platform in the world). We implemented the tool in three citizen research projects run by each of our cultural heritage partners (Royal Botanic Garden Edinburgh, Royal Museums Greenwich, The National Archives). We surveyed volunteers on these projects. We invited practitioners to three workshops that focussed respectively on

[1] We use the terms 'citizen research' and 'crowdsourcing' interchangeably in our work.
[2] https://www.nationalcollection.org.uk/about.
[3] Many cultural heritage organisations have increasing inclusion as a strategic aim, including for example The National Archives' *Archives for Everyone* <http://www.nationalarchives.gov.uk/archives-for-everyone>.

crowdsourcing and automation, the volunteer experience,[4] and the use of crowdsourced data by cultural heritage organisations.

Research methods

Our research approaches included desk research, an open call for information on cultural heritage crowdsourcing projects (which can be absent from the scholarly literature, focusing efforts instead on their practice), iterative tool development and implementation through three citizen research projects, community consultation including soliciting feedback on all aspects of the project by sharing progress throughout, and responding to guidance on best practice from our expert Advisory Board.

Covid-19 impact

Our work was affected by the pandemic, with members of the project team furloughed and taking on additional caring responsibilities. With people around the world confined to their homes, the numbers of contributors to citizen research projects increased significantly, and our projects are likely to have benefited from this surge. Some of the citizen research practitioners who shared their experience with us reported taking the lockdown as an opportunity to launch new and successful projects.

[4] We use the term 'volunteer' to refer to citizen participants in our crowdsourcing projects: to our knowledge all participants in our projects are volunteers, but volunteers work elsewhere in our projects too, in particular lending expertise to Royal Museum Greenwich's *HMS NHS: the Nautical Health Service* project, and to our Advisory Board.

Acknowledgements

This project is the result of collaboration. Project members have fed into all our project outputs. The deep thanks of us all go to colleagues across our organisations who fed into our work, to the members of our Advisory Board who donated their time and expertise to guide and reflect on our work, to the wider international community of colleagues working on citizen research and cultural heritage, including those who responded to our open calls and attended our workshops, and to the millions of citizen researchers who volunteer their time so generously to work on crowdsourcing projects, including our own three, and whose work and views informed this project.

Abstract

The *Engaging Crowds: citizen research and heritage data at scale* Towards a National Collection (TaNC) Foundation Project explored 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 was part of that wider landscape. The project was 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 had three focuses: community consultation on citizen research in cultural heritage organisations, including through workshops; prototype tool development for online crowdsourcing; and evaluating the tool through three citizen research projects and survey analysis. The project engaged with the wider community through seeking volunteers for the three citizen research projects and working with them once the projects launched; through our three workshops; through conferences; and through an open call for information about previous cultural heritage projects that used digitally enabled citizen participation. Taken together, the results of this work informed recommendations for best practice in encouraging and supporting meaningful public involvement with heritage collections. Our work feeds into the Towards a National Collection programme, enhancing our understanding of engaging the public digitally with cultural heritage.

Aims and Objectives

This collaborative project has helped us take steps towards a unified national collection by identifying and addressing current and future challenges facing the effective conduct, use, connection, and reuse of citizen research and the data it produces. Its objectives were as follows:

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 examining case-study projects we have analysed evidence on who participates in this work, what their motivations are, and developed our understanding of 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 participants can impact the results of that engagement. To help avoid bias in responses from one image to the next, some STEM projects,[5] for example, randomly serve images to participants. In cultural heritage contexts, and particularly for images of text, the ability for participants to navigate their own paths through content has been identified as key to maintaining interest in participation. Our project has developed an indexing tool that will enable this navigation. It has been evaluated through the three case-study projects on the Zooniverse platform, a volunteer survey and a workshop dedicated to the volunteer experience. The indexing tool has trialled this self-navigation method as a means of engaging participants more deeply in individual projects.

3. Explore barriers to and identify solutions for the effective use and reuse 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 using Artificial Intelligence (AI) and machine learning, and the public, including researchers and industry. Workshops including representatives of these audience groups have informed an exploration of the following questions:

- 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?

[5] STEM is a term used to encompass science, technology, engineering and mathematics. See: Science, technology, engineering, and mathematics - Wikipedia.

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

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

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

Partnership structure

The National Archives (TNA)

TNA led the project, its reports and surveys, and dissemination of project outputs. It created and supplied image data and metadata, provided records expertise, and ran a citizen research project on the Royal Hospital Chelsea's records. It also organised a workshop on automation in citizen research.

The principal investigator, project manager, and communications officer were based at TNA. A records specialist and citizen research project designer delivered TNA's citizen research project, with input from two research associates, 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)

RBGE created and supplied image data and metadata, provided records expertise, and ran a citizen research project on the plant family Gesneriaceae. It also organised a workshop about reusing citizen research data. Colleagues contributed to project reports and surveys, and to dissemination of project outputs. A co-investigator and project officer were based at RBGE.

Royal Museums Greenwich (RMG)

RMG created and supplied image data and metadata, provided records expertise, and ran a citizen research project on the Dreadnought Seamen's Hospital records. It also organised a workshop on the volunteer experience. Colleagues contributed to project reports and surveys, and to dissemination of project outputs. A co-investigator, research consultant, and super volunteer were based at RMG.

Zooniverse

Zooniverse has built, implemented and iterated the indexing tool for citizen research projects, trained the other project partners in use of Zooniverse tools (including the Project Builder and the newly-created indexing tool), helped with project setup, and provided expert review. Colleagues have contributed to project reports, surveys and dissemination of project outputs. A co-investigator, development manager, projects consultant and developer were based at Zooniverse.

Staffing structure

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

Chris Lintott, Professor of Astrophysics, Co-founder of Zooniverse, University of Oxford — Co-investigator. Led on indexing tool development, contributed to outputs and dissemination.

Elspeth Haston, Deputy Herbarium Curator, RBGE — Co-investigator. Led on RBGE's citizen research project and workshop on data reuse for collections-holding organisations, contributed to outputs and dissemination.

Martin Salmon, Archivist & Curator of Manuscripts, RMG — Co- investigator. Led on RMG's citizen research project design and delivery, and on volunteer experience workshop, contributed to outputs and dissemination.

Samantha Blickhan, Humanities Lead, Zooniverse — Development manager. Managed the development and implementation of the indexing tool, advised on citizen research project design and workflows, and community engagement. Contributed to outputs and dissemination.

Grant Miller, Communications Lead and Community Manager, Zooniverse — Projects consultant. Advised on citizen research projects including design and workflows, and community engagement. Contributed to outputs and dissemination.

Jim O'Donnell, Web Developer, Zooniverse — Indexing tool developer. Built indexing tool in three phases, and iterated based on feedback from wider project team.

Bernard Ogden, Research Software Engineer — Project designer. TNA citizen research project design, data workflow engineer for ingestion into TNA's catalogue. Analysed data from citizen research projects and created a data sharing platform.

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

Andrea Kocsis, Friends of The National Archives Research Fellow (Advanced Digital Methods), TNA — Research associate. Contributed 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.

Rebecca Hutcheon, Digital Scholarship Researcher - Research associate. Responsible for scoping the citizen research landscape and writing selected outputs (August 2021).

Ashleigh Hawkins, Digital Scholarship Researcher - Research associate. Responsible for scoping the citizen research landscape and writing selected outputs (February 2022 - April 2022).

Sally King, Digitisation Officer/Herbarium Volunteer Coordinator, RBGE — Project officer. Implemented RBGE citizen research project, contributed to volunteer engagement, organised a workshop on data reuse for collections-holding organisations.

Stuart Bligh, Head of Research and Information, RMG — Research consultant. Advised on RMG citizen research project and liaison across RMG (February 2020 - March 2021).

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

Louise Seaward, Head of Digital Research Programmes, TNA — Project manager. Managed all administrative aspects of the project, advised on workshop design.

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

Advisory board

Our Advisory Board met 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 brought valuable knowledge and experience from different perspectives around citizen research.

Adam Corsini, Collections Engagement Manager, Jewish Museum London

Stuart Dunn, Professor of Spatial Humanities, Department of Digital Humanities, King's College London

Libby Ellwood, Global Communications Manager, iDigBio

Siobhan Leachman, Citizen Scientist and Wikimedian

As part of her role on our advisory board, Siobhan Leachman summarised her thoughts in two reports which are available online (October 2020 and February 2022).

Revised overall programme

Date	Milestones	Work package
Dec 2019	TNA hosts first project 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	WP1 WP3
Apr 2020	RMG, TNA & RBGE begin citizen research projects design	WP5
May 2020	*Engaging Crowds* website goes live	WP6
Jun 2020	First call for information on citizen research landscape	WP2
Oct 2020	Zooniverse delivers prototype indexing tool Testing phase begins for RMG citizen research project Work on citizen research landscape report starts Project advisory board meets	WP3 WP5 WP2 WP1
Nov 2020	Testing phase begins for TNA citizen research project Zooniverse tests and refines indexing tool	WP5 WP3
Dec 2020	RBGE hosts second project workshop on data management and reuse Second call for information on citizen research landscape	WP2 WP2
April 2021	Testing phase begins for RBGE baseline citizen research project	WP5
Jun 2021	RMG citizen research project launches Zooniverse refines indexing tool for next project	WP5 WP3
Sep 2021	Second testing phase begins for TNA citizen research project	WP5
Nov 2021	TNA citizen research project launches Zooniverse refines indexing tool for next project Testing phase begins for RBGE indexed citizen research project 3 month no-cost extension granted	WP5 WP3 WP5 WP1

Dec 2021	RMG host third project workshop on volunteer experience	WP2
Jan 2022	RBGE citizen research project launches Second phase of RMG citizen research project launches Work on analysis of user engagement data begins Work on data sharing platform begins	WP5 WP5 WP5 WP4
Feb 2022	Second phase of RMG citizen research project launches (completion date is likely to be December 2022)	WP5
Feb 2022	Project advisory board meets	WP1
Feb 2022	Volunteer survey launched	WP2
Mar 2022	Citizen research landscape report completed	WP2
Apr 2022	Analysis of volunteer survey completed Analysis of user engagement data completed Data sharing platform launches Final report completed Project ends	WP2 WP5 WP4 WP6 WP1

Events and consultations

Date	Event	Subject	Number engaged	Notes
13 Dec 2019	People and Machines: co-creating with heritage collections. Workshop hosted by TNA (in person)	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. A post-workshop report is available in the Annex.
Jun 2020 onwards	Engaging with cultural heritage organisations and groups in a call for information	Requesting information on experiences of citizen research	30	These responses have been reviewed and written up as a survey of the citizen research landscape. See the Annex.
1 Dec 2020	After the crowds disperse: crowdsourced data rediscovered and researched Workshop hosted by RBGE (online)	Flow of data from citizen research projects back to collections-holding organisations, including quality control, ingestion and data reuse	60	The workshop discussed ideas on best practice 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 reused) effectively. A post-workshop report is available in the Annex.
2 Jul 2021	Cultural heritage hand in hand: how should we work with a community of citizen researchers? Workshop delivered by TNA at DCDC conference (online)	Surveying audience experience of working with online volunteers	30	This workshop aimed to gather insights from practitioners and researchers who have experience of supporting online volunteering. The findings of the workshop are summarised in the citizen research landscape report - see the Annex.

1 Dec 2021	Voices of the volunteers: the experience of online citizen research Workshop hosted by RMG (online)	Surveying how volunteers experience citizen research projects	32	The workshop allowed us to gain insights about the volunteer experience and also gather feedback on the Zooniverse indexing tool. A post-workshop report is available in the Annex.
Feb 2022	Volunteer survey	Surveying volunteers working on our citizen research projects	379	We asked volunteers about their experience of working on these projects and their views on the Zooniverse indexing tool. The survey results are summarised in the Annex.

Research approach/methods

Research and Discovery

We organised three workshops across the life of the project to gather information and ideas from the cultural heritage community. Our first workshop considered the potential and pitfalls of integrating automation and machine learning into citizen research projects. The second workshop brought together a broad audience of researchers and practitioners to explore the best ways of ensuring data generated in citizen research projects can be reused by institutions and the public. In our final workshop we invited volunteers to share their experience of participating in online cultural heritage projects and gathered feedback on the Zooniverse indexing tool. We held an open call for cultural heritage institutions to share their experience of citizen research with us, so we could reflect on their successes and challenges. We also ran a volunteer survey to further our understanding of the people who have participated in our projects. Findings from each of these tasks can be found in the Annex.

Finally, we drew upon the expertise of our advisory board in October 2020 and February 2022, who gave us recommendations for engaging volunteers and disseminating our results.

Indexing tool development

Development of the indexing tool took place in three stages, to coincide with the building and launch of the three citizen research projects. This approach allowed us to proceed through a complex development process in incremental steps, with built-in time for iteration at each phase based on user feedback as projects underwent beta testing and public launch. A blog post on the Zooniverse website gives a detailed overview of how the new infrastructure compares with traditional Zooniverse methods.

The first phase (for the *HMS NHS* project) included basic user selection options for workflow and subject set, with sequential classification within a given workflow (i.e. volunteers choose workflow and subject set, and are then shown pages in sequence while transcribing). The second phase (for *Scarlets and Blues*) built on the approach from phase one, but added the full indexing tool option, which included the ability to designate what metadata is shown to volunteers as an index, and allow volunteers to choose an individual subject to classify within a given project workflow. The final phase (for the *RBGE Herbarium* project) included all the work from previous phases, but added a previous/next pagination option within the classify page, so volunteers could essentially scroll through a dataset to decide what to work on, rather than having to go back to the index page.

This method allowed us to reconceptualise the underlying infrastructure for Zooniverse projects (a major undertaking) in a way that considered the technological implications (relationship to other projects on the platform, likelihood of bugs etc.) as well as user experience and volunteer feedback.

Citizen research projects

We used the Zooniverse Project Builder to create three new citizen research projects. Each project went through alpha and beta testing. Based on feedback from testers and the Zooniverse team, each project adjusted their workflows to make them as simple as possible whilst ensuring that they enabled volunteers to produce useful data.

The RMG project *HMS NHS: The Nautical Health Service* is based on the records of the Greenwich-based Dreadnought Seamen's Hospital, which cared for sick and injured merchant seafarers entering the port of London from 1826–1986. The project covers the years 1826–1930. Although highly structured with eighteen columns over a two-page spread, designing workflows that could deal with all the changes inherent in over a century of medical data was a challenge. In the final version, each column is presented as a separate workflow, such as Name, Date of Entry or Medical Complaint and volunteers transcribe from the top of the column to the bottom. The indexing tool allowed them to select individual registers to work on, with pages from the register then displayed in sequential order. This allowed volunteers to choose to focus on a particular year. *HMS NHS* is ongoing and 1,924 volunteers have taken part at time of writing.

In the TNA project *Scarlets and Blues*, volunteers transcribed records from the Special Board of the Royal Hospital Chelsea. The project was built on two workflows, one to transcribe meeting minutes (*Meetings)* and one to transcribe lists of names in the books' indexes (*People)*. Volunteers use the indexing tool to choose records in two stages – first selecting a period of time or a book and then picking a page, date, or "first letter of surname". Volunteers thus could work in particular periods, or follow cross-references. Workflows and instructions were simplified based on feedback but remained relatively complex. Over 500 *Scarlets and Blues* volunteers transcribed 2,000 pages from these minute books.

The RBGE project *The RBGE Herbarium: Exploring Gesneriaceae, the African violet family*, was based upon herbarium specimens of the Gesneriaceae family. The specimens are pressed plants mounted alongside a collection label which contains information on where, when and by whom the specimen was collected. The project had two workflows: in the *Latitude/Longitude* workflow, volunteers transcribed latitude and longitude data from the collection labels, and in the *Geography* workflow they transcribed country and lower geography and altitude information. The index grouped records by geographical region and allowed selection by botanist or scientific plant name, allowing volunteers to follow a person or plant-focused transcription path in different parts of the world. Specimens could also be paged through in order. 412 volunteers transcribed 3,568 of these collection labels.

Data sharing platform development

Giving volunteers access to the data they help to create is a vital part of citizen research. Discussions about best practice for data sharing were ongoing across the project, drawing in feedback from our advisory board and participants in workshops. We took time to ascertain the licensing conditions for each institution and reflect upon the best way to share images and data online. We created a data sharing platform on the project website for this purpose.

Research results

The project addressed the following research questions and in-depth reports are linked from the Annex.

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?**

We heard directly from our volunteer communities, via a workshop led by RMG and an online survey open to all participants in our citizen research projects. In the workshop, volunteers agreed a connection with the subject and contributing to something 'bigger' were important factors in their participation. All appreciated being able to contribute as little or as much as they could around busy lives, as well as the choice in how much they engaged with other volunteers. Above all, a choice of tasks that catered for different learning styles added appeal and deepened the sense of involvement.

Responses to our volunteer survey showed that online cultural heritage citizen research has an overwhelmingly positive impact on participants. Respondents cited a mixture of altruistic benefits, such as helping with research, giving back, and increasing access to records, and benefits to the individual, including providing opportunities for learning and development, filling in time, and gaining a sense of community.

We also looked at classification data relating to the number of volunteers and the timing of their activity. Our limited analysis indicated that different workflows can be associated with different patterns of engagement. Across the projects (and in citizen research generally) a small proportion of volunteers contribute the majority of the work, so perhaps we need to think about maximising the value of engagement for the majority of volunteers who will engage more briefly with a project.

2. **How does the ability to navigate one's own path through the data of a citizen research project affect engagement with the project?**

Based on feedback from beta testing, people overwhelmingly appreciated the ability to use the indexing tool to choose what they work on. All three projects saw classification rates that were within the normal range for transcription projects (i.e. people did not participate in these projects at a lower rate than on other Zooniverse projects not using the indexing tool).

Feedback in the volunteer survey also indicated that the majority of people appreciated working in new ways with the indexing tool. Several respondents stated that they liked being able to have a sense of control over what they were working on. Of those who did not like the indexing tool, most either experienced technical difficulties or were unaware of the tool. While some respondents were ambivalent about the tool or able to find both positives and negatives to its use, many respondents found it helpful, efficient and easy to use, and enjoyed being able to select material to work with.

The shape of the indexing tool was developed iteratively and now that it is fully formed there is the potential to apply different modules of it to different projects. For example, in hindsight the final iteration of the tool used in *The RBGE Herbarium* may have been well-suited to the transcription of admissions registers in *HMS NHS.* We do not currently have enough data to understand whether volunteers used the tool to follow a thread through the records but a more complete picture could be formed with future work. The indexing tool will be extended and studied further by Zooniverse in an upcoming project funded by the National Endowment for the Humanities.

3. How can we verify, assess, present, and value the contributions of citizen research?

In our study of the citizen research landscape and our workshop on data reuse we found that communication is vital both during and after a project. Volunteers, motivated as they are by a desire to contribute, who then develop active interest in the project, benefit most from regular feedback. Organisers, in short, need to build responsiveness into their workflows and maintain open channels of communication throughout the duration of the project. At the end of the project, it is vital that volunteers are recognised and kept up-to-date with the outcome of their work, and how the data is, will, and can be used.

Careful project design is key and critically must take into account any potential barriers to data access and use. Rigorous testing is required to ensure that the platform will be simple for volunteers to use and produces data that meets the needs of the organisation and that can be incorporated into its other systems. A more collaborative approach between volunteers and staff should be explored. Data quality concerns can be alleviated by the application of attribution to all data sets be they produced in-house by staff, an AI output or crowdsourced.

4. How can we enable the reuse 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, reuse and analysis?

Our second workshop focused on data reuse and discussions made clear that this needs to be considered in the design stage of any project. The ability to access data produced by volunteers was seen as a priority as part of a project's ethical responsibility. Integration of the data into Catalogue Management Systems (CMS) cannot always be achieved in a timely way: publishing raw data or using interim repositories were agreed to be acceptable stepping stones. Ideally project designs should aim for open access data in a variety of formats to suit the needs of different audience demographics. Responsibilities for each stage of data creation and movement should be clearly identified and institutional gatekeeping of the process avoided.

In our first workshop on machine learning we explored the possibility of using crowdsourced data to train algorithms, such as for handwritten text recognition. Datasets created as a result of a crowdsourcing project using machine learning will require additional metadata, for example, to document the model and training data used. Transparency around decision-making, the development/selection, and training of machine learning models and tools is necessary in order to enable their reuse, and to ensure transparency and openness with volunteers.

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?**

Our advisory board stressed the importance of making licensing and citation information explicit to encourage volunteers to understand how data can be reused. Data created through citizen research should be shared as soon as possible. This can be in raw form in an intermediate platform if ingestion into the institutional CMS is not straightforward. Raw data, and even a collection of processed transcriptions, require more effort and skills to engage with than data presented through a CMS. Providing multiple output formats that are compatible with different systems and use cases will broaden accessibility and reuse. This could be a searchable reader-friendly book format and a raw csv file for more quantitative research as exemplified in the *Mutual Muses* project. Building interface(s) for particular lenses on the data will also broaden access. Training in data access and use should be provided by institutions for users both online and face-to-face to facilitate access and use of the data.

Project outputs

Citizen research projects

HMS NHS: The Nautical Health Service (ongoing)

This project, led by RMG, is based on the records of the Greenwich-based Dreadnought Seamen's Hospital, which cared for sick and injured merchant seafarers entering the port of London from 1826–1986. The first part of the project, which focused on the years 1826–1930, launched in June 2021. 1,600 volunteers completed the transcription of 49,000 images in this first phase of the project. The second phase of the project was launched in February 2022, with transcription continuing at the time of writing.

Scarlets and Blues (complete)

This project, led by TNA, launched in November 2021 and was completed in January 2022. It drew upon five minute books of the Special Board of the Royal Hospital Chelsea from 1908-1919.

The RBGE Herbarium: Exploring Gesneriaceae, the African violet family (complete)

RBGE's project launched in January 2022. It involved the transcription of selected elements of the herbarium collection label data from Gesneriaceae herbarium specimens. The project ran from January to February 2022.

Tools and platforms

Indexing tool

The indexing tool can be seen in each of the three citizen research projects, with varying degrees of 'completeness'. It allows volunteers to choose their own pathway through a project, rather than working on pages presented at random. Code relating to the tool can be found on the Zooniverse GitHub.

Data sharing platform

We have created a dedicated space on our project website to share images and classification data from our three citizen research projects. This page includes full licensing and citation information, to make it as easy as possible for anyone to reuse material from the project.

Reports

Citizen research landscape

This report[6] summarises responses received to our call for information about citizen research projects in cultural heritage institutions. It explores the successes of these projects and some of the challenges they encountered.

Workshops

- People and Machines[7], December 2019
- After the crowds disperse[8], December 2020
- Voices of the Volunteers[9], December 2021

These reports summarise discussions at the three project workshops - on machine learning, data reuse and the volunteer experience.

User engagement analysis

We analysed volunteer activity on each of the three citizen research projects to gain an insight into user behaviour. This analysis[10] focuses on questions around time of volunteer contribution and the number of active volunteers on each project.

Volunteer survey

In March 2022, we circulated a survey to participants in our three citizen research projects. 379 people took part in the survey. The survey results[11] indicated that we had many active volunteers who valued the opportunity to make a contribution to citizen research. Feedback on the indexing tool was generally positive, with some volunteers being unsure about its function but others appreciating the increased agency compared with other Zooniverse projects.

[6] https://doi.org/10.5281/zenodo.7079083
[7] https://doi.org/10.5281/zenodo.7079535
[8] https://doi.org/10.5281/zenodo.7081409
[9] https://doi.org/10.5281/zenodo.7151964
[10] https://doi.org/10.5281/zenodo.7151974
[11] https://doi.org/10.5281/zenodo.7151994

Dissemination

Conference presentations

We have presented at the following conferences:

- **5th AHRC Connected Communities Heritage Network Symposium, February 2021.** Pip Willcox and Louise Seaward, 'Connecting with the crowd: creating and supporting citizen research online'.
- **Discovering Collections, Discovering Communities conference, July 2021.** Pip Willcox, Bernard Ogden and Louise Seaward, workshop on 'Cultural heritage hand in hand: how to work with a community of citizen researchers'.
- **DARIAH Annual event 2021: Interfaces, September 2021.** Grant Miller and Sam Blickhan, part of a panel on ‘The Interface(s) of a Virtual National Collection’.
- **International Congress on Archives: October 2021.** Pip Willcox, ‘Empowered people, empowering society: citizen research and shared heritage’.
- Doing Maritime History Research Online, British Commission for Maritime History, February 2022. Martin Salmon, ‘Travelling the Zooniverse: Medical Data from the Dreadnought Seamen’s Hospital’.

Blog posts

Date	Partner responsible	Topic
July 2020	TNA	Introduction to Engaging Crowds
Sept 2020	RMG	Introduction to citizen research project
Oct 2020	RBGE	Invitation to attend RBGE workshop
Mar 2021	RBGE	Report on RBGE workshop
Jun 2021	RMG	Launch of RMG citizen research project
Nov 2021	RMG	Invitation to attend RMG Workshop
Nov 2021	Zooniverse	Overview of indexing tool
Nov 2021	TNA	Research relating to TNA citizen research project
Dec 2021	TNA	Research relating to TNA citizen research project
Jan 2022	RBGE	Launch of RBGE citizen research project
Apr 2022	TNA	Insights from developing TNA’s citizen research project

Recommendations for the programme

Through this project, we have explored best practice in citizen research, the experience of volunteers, the challenges and potential facing practitioners in the field. Our findings can help to enhance the efficacy and impact of citizen research in the cultural heritage sector, foregrounding its ethics, and paving the way for a future national and supranational citizen research effort.

Our recommendations can be summarised under the following headings. For a fuller discussion of the issues covered here, please see the Annex.

Setting up a citizen research project

Citizen research projects are usually set up to achieve a specific short-term goal that an organisation may not be able to realise without the help of volunteer effort. The resources required to set up and run such a project should not be under-estimated.

Project teams need to think about the format and quality of data that would be most useful and make sure their project is designed to support volunteers to produce this. Before a project is launched, teams should know what data they will be receiving from volunteers and have mechanisms in place to process and share this publicly.

Task variety and flexibility is important for encouraging a wide range of people to take part; some people may be drawn to certain tasks but inclined to avoid others. Projects with short, simple tasks that can be fitted into different moments of the day have the best chance of engaging a large number of volunteers. Setting up simple tasks can be challenging in the cultural heritage field where records are often ambiguous and complicated. Projects need to be prepared to provide more support to volunteers, or deal with increasingly messy data from volunteers who may have forgotten or misinterpreted complex instructions. Organisations need to find a balance between their needs and the volunteer experience, and be open to surprises as volunteers respond to the collections.

Projects should be rigorously tested and iterated before launch and project teams should be ready for the possibility that they will need to reshape projects radically based on feedback. Even after volunteers are on board, the best projects do not remain fixed from the outset but rather continue to be adjusted and updated based on the experience of volunteers and staff.

Organisations should consider how much support they can offer to volunteers during the lifetime of a project. The less support per volunteer (necessarily quite low in very large-scale projects), the more intuitive the workflows must be.

Long-term sustainability of citizen research projects and their outputs should be planned from the outset. How long will the project remain active and are there staff available to support this? How will the outputs of the project be made publicly available and what resources are needed to accomplish this?

For all these aspects, projects need support from across organisational structures. Senior sponsorship can be helpful in facilitating this across the lifetime of a project. We would recommend publishing best practice for crowdsourcing in cultural heritage, and note the publication of a useful handbook in this area.[12]

Licensing and copyright in citizen research projects

Questions around the digitisation of collections were out of scope for this project, and we acknowledge that this is a prerequisite for many of the types of project we explored. This requires resources and how this is funded may influence its licensing for reuse.

Organisations should start to discuss licensing and copyright requirements in citizen research projects as early as possible in the planning stage. Arrangements will need to be made for the collection, use and later reuse of digitised images and data created by volunteers. The situation may be complicated in collaborative projects where organisations hold different positions on copyright. There may be challenges to navigate around image rights where organisations use these to generate income. Where projects include machine learning models, intellectual property rights and transparency should be carefully considered.

Citizen research projects enable volunteers to make a huge contribution to the cultural heritage field and this contribution must be clearly acknowledged in a prominent place on project websites. The ethical and legal requirements to protect any potentially identifying information from volunteers should also be factored in.

Licensing information about how images and data can be reused should be communicated as simply as possible, with any technical terms clarified and links to further explanations. If there are restrictions on the reuse of data produced by the project, these should be made clear to volunteers at the project launch as it may affect their choice to participate.

Examples of how to cite the project itself and data generated in the project should be provided. Further information about how to download and use data will empower those with less technical expertise to access it. The work of Towards a National Collection on licensing and copyright will feed usefully into citizen research projects.

Integrating automation into citizen research projects

With technology advancing rapidly, the question of how far to integrate automated approaches into citizen research is an urgent one. Machine learning could perform tasks perceived as tedious, freeing volunteers to channel their skills into more interesting challenges that machines cannot (yet) perform. With informed consent, data produced by volunteers could be used to train machine learning models.

[12] Mia Ridge, Samantha Blickhan, Meghan Ferriter and others, *The Collective Wisdom Handbook: perspectives on crowdsourcing in cultural heritage* (online, 2021) <https://doi.org/10.21428/a5d7554f.1b80974b>. This co-authored publication was the result of a book sprint organised by the *Collective Wisdom* project, funded through the National Endowment for the Humanities in the US and the Arts and Humanities Research Council in the UK, and involved two *Engaging Crowds* project members.

We need to consider further whether it is ethical to ask volunteers to complete tasks that could be undertaken by machines, acknowledging that some volunteers find enjoyment in simple tasks. There are complex considerations when it comes to working with machine learning. Only certain types of data and tasks are appropriate. Many organisations do not have the technical infrastructure and knowledge to support this kind of work and new tools will not fit with every project. Project tasks should be engaging for volunteers, whose motivations vary, and the requirements or potential for machines should not be prioritised over those of humans.

If organisations create new kinds of citizen research projects involving automation, there must be complete transparency about the technology. What models are being used and on what parameters? How has potential bias in the training data been addressed? Why are volunteers still needed if machine learning is available? How are tasks split between humans and machines, and why? This must all be explained in a jargon-free manner that it is accessible to a wide audience, informed consent can be given by volunteers, and ideally they have the possibility of opting out.

Best practice dictates that machine learning models should be shared publicly, with documentation on model generation, data sets and applications to support it. We recommend a shared platform for machine learning training data sets, models, documentation and tooling for use and reuse across cultural heritage including and beyond citizen research projects.

Increasing volunteer engagement in citizen research projects

Volunteers are motivated to participate in citizen research by a range of factors and the relative importance of these vary for each individual.

An interest in the subject matter and an affinity with the project mission is shared by many volunteers. Project teams should make sure that documentation clearly demonstrates the purpose and usefulness of the project. Volunteers value regular communication from project teams as this can foster a sense of community, receiving feedback on their work and understanding why their contribution matters. This communication should be two-way, so that volunteers can share their thoughts on the project and suggest improvements.

Projects should include a forum or similar space that allows volunteers to communicate with each other. Our findings indicated that many volunteers enjoyed talking to others, or at least having the opportunity to do so.

Although generating new data for an organisation is usually the rationale for crowdsourcing projects, valuing the work and time of the volunteers is equally vital. A balance may need to be struck between producing data in a format that is institutionally useful and creating simple tasks that volunteers will be motivated to complete.

Investigating volunteer activity can shed light on their engagement with a project, but statistics and data visualisations are slippery and can be misleading. When using data in this way, project teams should allow sufficient time and expertise to interrogate it.

Citizen research projects are as much about outreach and participation as they are about data creation. Project teams should consider how they can have a positive impact on volunteers rather than simply relying on their labour, such as by providing them with opportunities to develop new skills, share their expertise, have agency (as in the case of the Zooniverse indexing tool) or have fun. Here, as throughout our work, we recommend centring ethics in our collective practice.

Working with data generated by volunteers

Best practice in citizen research is to make volunteer outputs and source materials used in the project freely available for use and reuse, in multiple formats, as early as possible.

To ensure there will be a pipeline for ingesting volunteer data into institutional systems and make it public, project teams need to involve cataloguing and other colleagues across their organisations, from the project design phase onwards. In practice, this situation can be complicated. For many organisations, Collections Management Systems (CMS) are difficult to work with and access, sometimes under-resourced and with limited ability to include data created by volunteers. If full integration in institutional systems is not possible, projects should make sure they have alternative solutions in place from the start, such as sharing data through a data repository.

We recommend that licensing information is clear and visible as this is vital to facilitate reuse of data created through citizen research. Organisations should also make active efforts to encourage data reuse, such as through direct promotion or data tutorials. A programme-wide approach to sharing skills and publicising learning

Contacts

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

pip.willcox@nationalarchives.gov.uk

Samantha Blickhan, Humanities Lead, Zooniverse

samantha@zooniverse.org

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

louise.seaward@nationalarchives.gov.uk

Annexes and links

Project website

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

GitHub repositories

Zooniverse repository

HMS NHS aggregation code (still in development)

Scarlets & Blues aggregation code (still in development)

User engagement analysis code

Reports

- Citizen research landscape report https://doi.org/10.5281/zenodo.7079083
- People and Machines workshop report https://doi.org/10.5281/zenodo.7079083
- After the Crowds Disperse workshop report https://doi.org/10.5281/zenodo.7079083
- Voices of the Volunteers workshop report
- https://doi.org/10.5281/zenodo.7151964
- User engagement analysis report and figures
- https://doi.org/10.5281/zenodo.7151974
- Volunteer survey report
- https://doi.org/10.5281/zenodo.7151994

Summary of work on the indexing tool

This project allowed Zooniverse to completely rethink their approach to subject delivery and classification. The development effort and results for each phase is described in detail below.

Phase 1: *HMS NHS: The Nautical Health Service*

This work included:

- Creating infrastructure that allows volunteers to choose what workflow and subject set they want to work on and to classify in sequence, as part of the Project Builder (Zooniverse's free-to-use crowdsourcing platform).
- Creating banners that show volunteers what 'page' they are on within a given subject set.

- Creating a ‘simple’ dropdown menu in our new front-end infrastructure.
- Displaying workflow and subject set completeness (shown as %) on the homepage and subject set selection modal respectively. Completed workflows are removed from the homepage; completed subject sets are greyed out and moved to the end.
- Adding administrative tools to the Project Builder for implementing subject set selection and sequential classification at a per-workflow level, supporting wider reuse.
- Building out existing ‘copy workflow’ capability so that administrative-level settings (like the indexing tool) persist in copied workflows, supporting wider reuse.
- Fixing bugs around session storage and repeat subject delivery.

Phase 2: *Scarlets and Blues*

Scarlets and Blues allowed Zooniverse to expand upon the tools created for *HMS NHS*. This work included:

- Creating an index modal based on subject metadata. Involved creating a new API endpoint: /subjects/selection and pulling in metadata from subject manifest (.csv) fields that begin with the % symbol.
- Creating Slack commands that kick off the process of building the index. Once the subject sets are uploaded to the Project Builder, any Zooniverse team member can start or update the index build via Slack command.
- Creating a ‘completed’ modal that will block the task area for subjects that have been fully classified. This is necessary to keep volunteers from working on subjects after retirement unless they opt in (this can be helpful for testing or classroom use). The modal allows volunteers to choose to classify a subject anyway, or to jump to the next available subject for classification. If a subject set is fully complete, they can go back to the set selection page and choose a new one to work on.
- Adding support for recursive, branched workflows and persistent annotations (necessary for complex workflow types)
- Displaying workflow, subject set, subject IDs in project URLs. This helps to ensure that the page always reloads the active subject ID.
- Updating sequential classification service from Cellect to Designator.

Phase 3: *The RBGE Herbarium: Exploring Gesneriaceae, the African violet family*

The *RBGE Herbarium* included all features created for *HMS NHS*, with additional efforts including:

- Creating the ability for volunteers to ‘paginate’ through a subject set from the classify screen. These previous/next buttons allow volunteers to scroll through the set until they see an image they want to work on, and fixes a long-standing expectation of being able to load a new subject (typically via the Refresh Page option) which was rendered unusable due to the sequential subject delivery methods used here.

Advisory board recommendations

Our Advisory Board met twice during the project to guide our work and dissemination. The following summarises their recommendations and our actions:

Board emphasised the importance of volunteers being able to reuse and share content they helped to create. Datasets should be shared with clearly explained licensing details, citation information and a link to any code we have developed.

Our response:

- Institutions have spent time working with their licensing teams to understand the conditions by which we can share data and images and to include clear explanations for volunteers to consider before beginning work.
- Data produced by volunteers and links to images is shared on a dedicated page on the *Engaging Crowds* project website. This page has full details of licensing conditions, citation information and code.

Board advised that we could link up with local groups, digital clubs or university classes to attract a broader range of volunteers.

- Due to pandemic-related delays, our three citizen research projects launched later than planned. Because of this, we did not have capacity to organise wider promotion to volunteers. Despite this, our projects were able to reach hundreds of volunteers.

Board advised that existing volunteers from our institutions may not appreciate transcribing in a new system.

- Our citizen research projects were primarily promoted via social media and via the Zooniverse newsletter, with the aim of attracting new volunteers to our institutions.
- Existing volunteers from our institutions were invited to take part in our projects. We created communications around this which emphasised that these projects were part of an experimental research project.

As part of the project, we produced a report on the citizen research landscape in the UK. Board advised that small organisations might find this report helpful if it explained the steps to build a citizen research project.

- This is out of scope for the report, which focuses on lessons learned from existing citizen research projects.
- We have included reflections on the process of building our individual projects across the final report.

- We have signposted other relevant resources in the report, such as the Zooniverse Project Builder and the recently produced handbook by the AHRC-NEH-funded *Collective Wisdom* project. A Zooniverse team member was co-investigator of the project, and a team member from TNA participated in the book sprint that co-authored the handbook.

Board advised that the citizen research report could include activity of informal or volunteer-run organisations.

- We sent out an open call for responses for this report, across various mailing lists, social media and live events with different audiences.
- All of the responses were from larger organisations, who may have had more capacity to write up their findings from such projects.
- The work of these small-scale Foundation Projects is one part of the Towards a National Collection programme. A large-scale Discovery project, *Our Heritage, Our Stories*, with TNA as the lead cultural heritage partner, is focusing community-generated digital content including crowdsourcing projects, and will take this work further with the greater capacity the larger project enables.

As part of the project, we circulated a survey to volunteers working on our projects. Board advised that this survey should include qualitative questions to help gather information about deeper volunteer engagement.

- Our survey included qualitative questions and space for respondents to add any further comments. A summary of the findings of the survey are included in the final project report.

Board advised that the final project report should contain best practice recommendations, be disseminated in shorter and longer forms and be available open access.

- The final project report and additional annexed material are freely available via the Towards a National Collection website.

Board advised that we archive each of the citizen research projects, so that they can be studied alongside the data in the future.

- We explored various options for archiving our citizen research projects including Browsertrix and the Internet Archive Wayback Machine.
- We found that the Internet Archive Wayback Machine could only archive an incomplete version of each project, with limited interactivity. Browsertrix captured the sites more fully but errors remained around missing pages and images.
- The most feasible and fast option for each of our institutions was to create videos detailing different pathways through our projects.
- Both workflows in *Scarlets and Blues* and *The RBGE Herbarium* have been archived in this way. As the first phase of *HMS NHS* had a greater number of workflows, one workflow from each

workflow 'type' has been archived. The second phase of *HMS NHS* will be archived in the same way once it has been completed by volunteers.

- The videos which act as an archival record of our project will be available on the *Engaging Crowds* project website.

Board advised that we promote our work as much as possible, so that the cultural heritage community can learn that challenges around working with citizen research data can be navigated.

- Once the final report is released, we will promote it via our institutional channels.
- We will continue to participate in dissemination via the Towards a National Collection programme.