TOWARDS A NATIONAL COLLECTION | UKRI Arts and Humanities Research Council

FOUNDATION PROJECTS

SUPPORTING DOCUMENTATION

After the Crowds Disperse

Engaging Crowds: citizen research and cultural heritage data at scale

Elspeth Haston, Sally King
Royal Botanic Garden Edinburgh

Contents

Introduction

Over the past decade, cultural heritage organisations have increasingly turned to volunteers through citizen research (or crowdsourcing) platforms to help make their holdings digitally accessible for discovery and research. These platforms have proved to be highly successful, both in attracting and retaining the interest of volunteers, and in the rate and quality of data collected. They have provided a new and deeper level of engagement and attracted more diverse audiences to national collections. However, there are still significant hurdles to overcome to achieve seamless sharing and movement of data between institutional Collection Management Systems (CMS) and crowdsourcing platforms, and back again. If crowdsourcing is to fulfil the potential of its generous volunteers this break in the data cycle must be closed.

Scope

This workshop was part of a series of workshops within the project Engaging Crowds: citizen research and heritage data at scale: a collaboration between The National Archives, the Royal Botanic Garden Edinburgh, Royal Museums Greenwich and Zooniverse at the University of Oxford. Engaging Crowds is a Foundation Project within the Arts and Humanities Research Council-funded Towards a National Collection programme.

This workshop was led by Royal Botanic Garden Edinburgh and took place online in December 2020.

This interdisciplinary workshop explored the use of citizen research projects to promote the existence of and access to collections, methods of quality control and analysis to ensure crowdsourced data can perpetually be reused, and the current practical barriers to data ingestion into collection management and discovery systems.

In order to examine these questions, we put out a call for expressions of interest from cultural heritage practitioners and researchers from a range of disciplines including historians, curators and collections specialists, researchers using collections, web designers, CMS developers, citizen research platform developers, data scientists, Human-Computer Interaction researchers, social scientists working within heritage, academic, industrial, or third sector disciplines. There were a total of 54 participants, including the organisers.

The workshop was structured in two halves. The first focused on the potential barriers and practical implications for data re-ingestion and reuse through a series of presentations from across the cultural heritage sector. In the second half of the workshop participants worked in small groups to consider questions that explored how to close the gap between data extraction and rediscovery.

This report recounts the presentations and discussions at the event, along with a set of recommendations that were identified to promote data reuse in crowdsourcing projects.

Summary of key recommendations

Project design should consider the purpose of data collection and the data's destination and ingestion pipeline.

A collaborative relationship between CMS managers and crowdsourcing project leads promotes closer alignment of data schemas.

Crowdsourced data should be open access and available in a variety of formats to facilitate compatibility with different systems and use cases.

Where ingestion is not possible, implement interim data-sharing solutions, such as making the raw data publicly available in a searchable format.

Add Digital Object Identifiers (DOIs), attribution information, citation requirements and instructions on access to all datasets to increase visibility and confidence in crowdsourced data.

Data usage rights should be clearly communicated to all stakeholders at the project outset.

Avoid rigidly defined roles of data contributor and data consumer to prevent an imbalance in project benefits.

Use a variety of media opportunities to diversify engagement with crowdsourced data.

Workshop participants

Organisers: Sally King (Royal Botanic Garden Edinburgh), Elspeth Haston (Royal Botanic Garden Edinburgh), Robyn Drinkwater (Royal Botanic Garden Edinburgh) and Louise Seaward (The National Archives).

Presenters: Pip Willcox (The National Archives), Melissa Gill (J. Paul Getty Trust), Nathaniel Deines (J. Paul Getty Trust), Tom Blake (Boston Public Library), Samantha Blickhan (Zooniverse / Adler Planetarium), Mike Denslow (University of Florida), Daniel Pett (Fitzwilliam Museum) and Chiara Bonacchi (University of Stirling).

Presentations summary

Pip Willcox presented an overview of the *Engaging Crowds* project. A summary of guest presentations follows.

Data Re-use

Mutual Muses by Melissa Gill and Nathaniel Deines (J. Paul Getty Trust)

The *Mutual Muses* project sought to transcribe the correspondence between Lawrence Alloway and Sylvia Sleigh, key figures of the twentieth-century art world. The focus was to create quick access to open data (Creative Commons licence CC0) for two main audiences – the volunteer transcribers and digital humanities scholars. Through consultation with these user groups an interim solution was developed in the form of a GitHub repository. This platform allowed the project data to be shared as both a chronologically readable book and a research-friendly dataset. The project's long-term goal was to host the data on their institutional collections viewer.

Data Quality

Anti-Slavery Manuscripts by Tom Blake (Boston Public Library) and Samantha Blickhan (Zooniverse / Adler Planetarium)

The *Anti-Slavery Manuscripts* project (ASM) was one of two bespoke text transcription projects that formed part of the Zooniverse-led Transforming Libraries and Archives Through Crowdsourcing project, funded by a National Leadership Grant from the US institution for Museum and Library Services. The research aims of this project included exploring whether the then-current Zooniverse methodology of multiple independent transcribers produced better results than allowing volunteers to collaborate. As part of the ASM project, the team ran an A/B experiment looking at the quality of individual versus collaborative transcription methods on the same dataset.

The experiment found that volunteers who were able to see and interact with other people's transcriptions produced significantly higher-quality data with a quicker turnaround time.[1] The results from the ASM project have been reviewed by the Boston Public Library team using the newly-developed Aggregate Line Inspector and Collaborative Editor (ALICE) tool, developed by Zooniverse with funding from the US National Endowment for the Humanities. ALICE launched publicly in 2021 as a free-to-use tool for researchers building text transcription projects on the Zooniverse platform. The tool allows volunteers to work collaboratively to view, annotate and

[1] Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, Amy Boyer, Andrea Simenstad, *et al.* 'Individual vs. Collaborative Methods of Crowdsourced Transcription'. *Journal of Data Mining and Digital Humanities* (2019), <https://hal.archives-ouvertes.fr/hal-02280013v2> [Accessed 16 March 2021].

edit the output of text transcription projects. The example of ASM shows how project design can help to ensure that volunteers produce data of the highest possible quality.

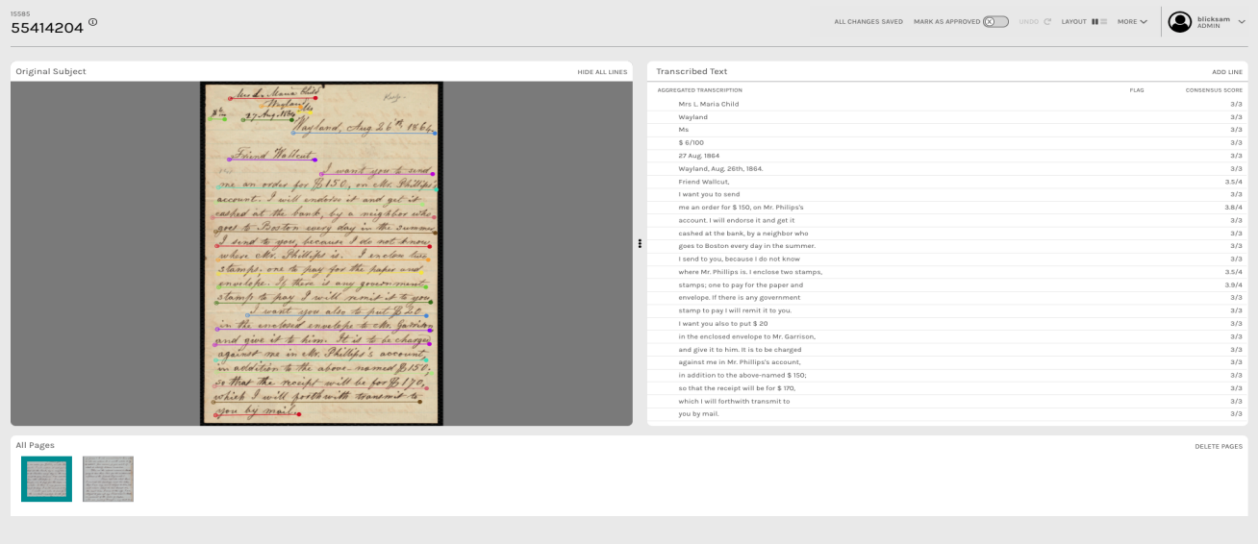

Figure 1: A screenshot of the ALICE interface,with an annotated image of the letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864 (left) and corresponding transcriptions of each line of the letter (right). [2]

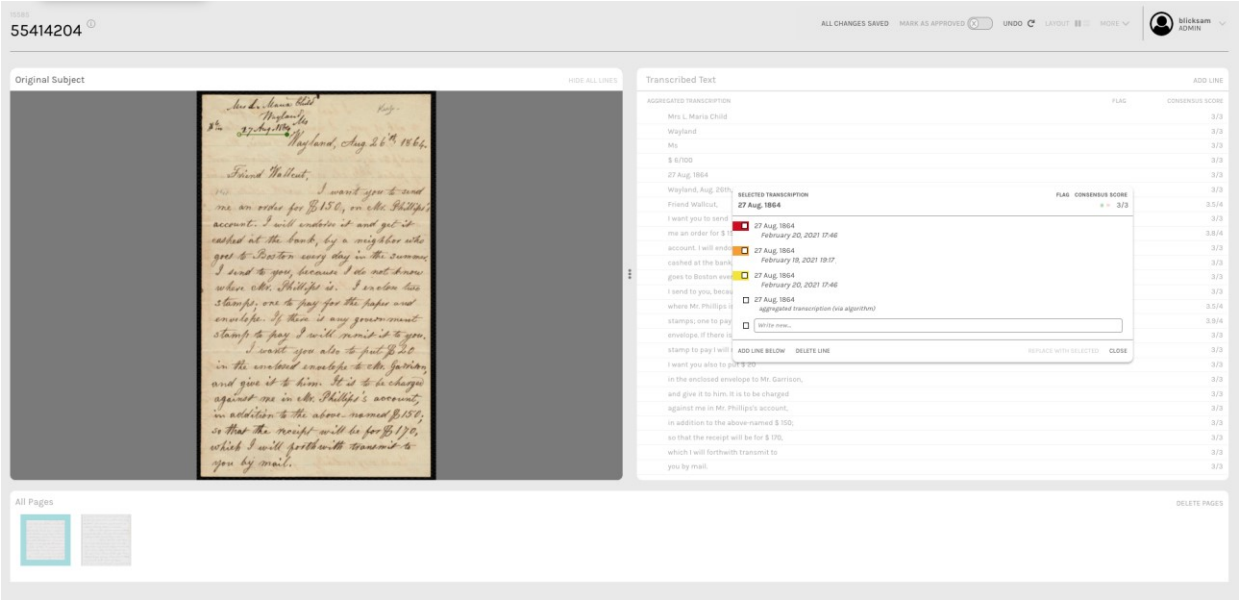

Figure 2: A screenshot of the ALICE tool showing an image of a letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864 (left) and three transcriptions carried out by three different volunteers [usernames have been redacted] and an aggregate of the three (right). Note the consensus score (top right of the selected transcription pop-up window, right) which indicates the number of transcriptions and the number of these that agree.[3]

[2] The Zooniverse, A screenshot of the ALICE interface with an annotated image of the letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864, and corresponding transcriptions of each line of the letter, screenshot, Zooniverse <https://alice.zooniverse.org/> [accessed 23 March 2021].

[3] The Zooniverse, A screenshot of the ALICE interface showing an image of a letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864 and three transcriptions carried out by three different

Data ingestion and research

Notes from Nature by Mike Denslow (University of Florida)

Notes from Nature, on the Zooniverse platform, is a community-focused project that seeks to engage the public in transcribing collection labels of natural history specimens. So far, the project has captured over 3.5 million transcriptions from the estimated 10 billion specimens housed in Natural History collections across the globe.

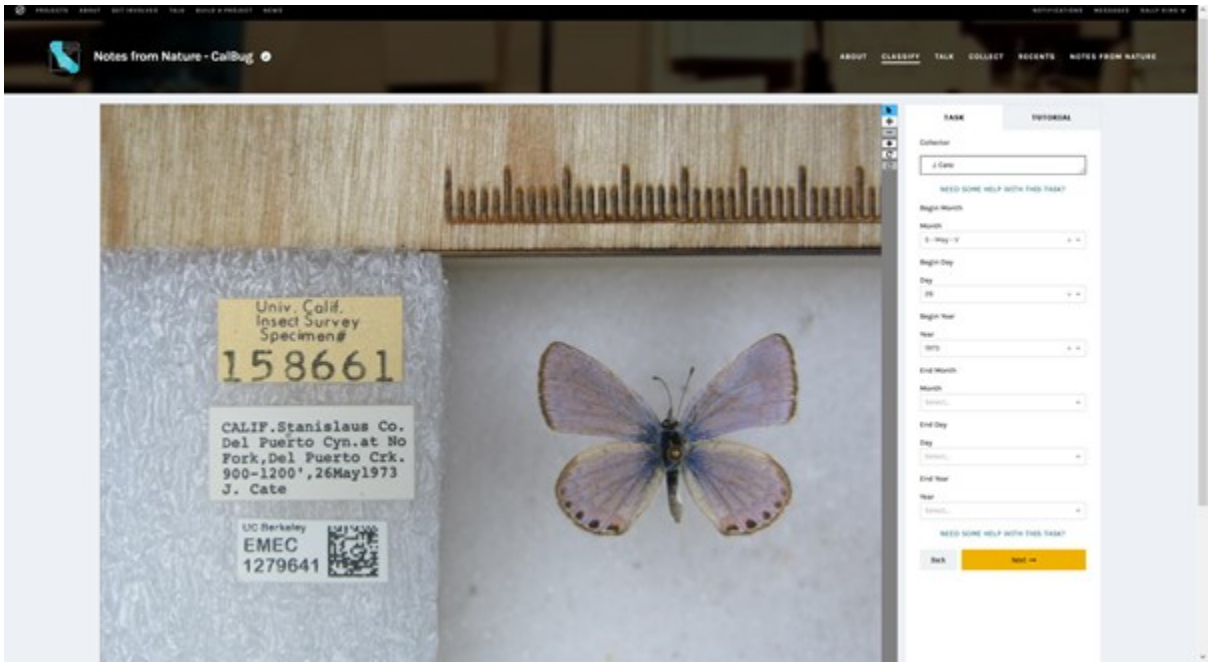

Figure 3: An example of a transcription form, from one of the many Notes from Nature projects hosted on Zooniverse. This project 'Calbug' aims to transcribe specimens from the California Terrestrial Arthropods Database.[4]

This project has an established ingestion protocol which allows transcription data to be added to pre-existing stub records in the South East Regional Network of Expertise and Collections (SERNEC) data portal (one of 18 portals within the Symbiota family that house over 18 million biodiversity records). This project is an example of a strong relationship between a large content discovery platform and a public engagement project. Such a relationship enables accessible, traceable data that is research-ready.

volunteers [usernames have been redacted] and an aggregate of the three, screenshot, The Zooniverse <https://alice.zooniverse.org/> [accessed 23 March 2021].

[4] UC Berkeley, *UC Berkeley EMEC 1279641*, screenshot, The Zooniverse <https://www.zooniverse.org/projects/md68135/notes-from-nature-calbug> [accessed March 2021].

Data re-use in research

Micropasts by Daniel Pett (Fitzwilliam Museum) and Chiara Bonacchi (University of Stirling)

Micropasts is centred around three open pillars: open data, open software and open coding. Regardless of the stage and progress of the project, the ethos is to make the data instantly available. It takes participation to the next level, creating a collaborative space between academics and other communities in archaeology.

Figure 4: Sketchfab, Micropasts crowd-produced annotated 3D model of a Bronze age spearhead from Arreton Down, Isle of Wight, UK. Fiske, Hugh via the crowd. Licensed under CC BY 4.0.[5]

The result is co-produced open data sets of a variety of media. These can then be used in co-designed follow-up research by the same groups using open source software, Pybossa, and additional reproducible, shareable code.

[5] Fiske, Hugh via the crowd. *Bronze Age Spreadhead SL. 746.,* screenshot, Sketchfab, 2021, <https://sketchfab.com/3d-models/bronze-age-spearhead-sl746-2e87e13023774a27ac8f59c9a231cfcf > [accessed 23 March 2021].

Discussions summary

Participants were split into groups and offered one of three questions to explore. The following summarises these discussions and pulls together challenges and recommendations under each theme.

Reingestion

How do we easily ingest and integrate citizen research (or crowdsourced) data into Collection Management Systems (CMS)?

Challenges

Crowdsourced data is often destined for content management systems (CMS). Mapping the crowdsourced data to these systems can be highly problematic as many CMS do not have the capability to hold transcription data and associated metadata. For one institution where ingestion was possible, the size of transcription data impacted the behaviour of their discovery platform, causing an uneven weighting within catalogue search results. This was due to objects with associated transcription data having a disproportionate amount of searchable metadata in comparison to other collection objects.

Another perceived barrier to ingestion is the quality of crowdsourced data. It is commonly thought a large amount of time and resources are required to clean it to an acceptable standard for ingestion. This is compounded by institutions not consistently having the in-house skills for data aggregation and standardisation to 'massage' the data into CMS.

The involvement of multiple stakeholders with different interests and priorities adds further complication to post-project data management.

Recommendations

Prior to the writing of a proposal for a crowdsourced project the whole project design should be clearly defined and include a data ingestion and management plan. This should be informed by looking at the project purpose: is it engagement with collections, data capture or data enhancement? The destination for any data needs to be resolved: will it be associated with an object image or an object data record? The person(s) responsible for the data at each stage (generation, ingestion, discovery, reuse tracking and preservation) need to be identified, keeping in mind the need and feasibility both in terms of staff time and cost, for sustained management of the data.

Where possible, working together with CMS developers at the planning stage would help to create an effective flow of data and solidify a culture where crowdsourcing projects are seen as

beneficial to all stakeholders. In many instances, however, this interaction is not possible. It also assumes that the CMS is regularly updated to meet individual organisations' needs. While some institutions have bespoke in-house CMS, many use commercial 'off the shelf' systems that aim to cater for a broad range of institutes with a limited degree of customisation.

In cases where CMS ingestion is not possible, interim solutions can be put in place to make crowdsourced data discoverable and searchable outside a CMS. This might include storing it in a GitHub repository: providing both researcher-friendly CSV format for access at scale, and a readable 'book format' for readers interested in individual details. However, while GitHub has 'brand recognition' within the sciences it appears not to be as well-known and therefore accessible to other disciplines and sectors, including many cultural heritage collections.

Another solution is the implementation of an adjacent database associated with the CMS data records by the object's unique identifier. The data can then be hosted on the institution's online discovery service alongside each CMS record in a searchable format. It is noted that this solution, which has been implemented at Royal Botanic Garden Edinburgh requires a developer and staff with the knowledge to import data into the database and ensure the service is maintained. It is unlikely that smaller institutions will have access to these resources.

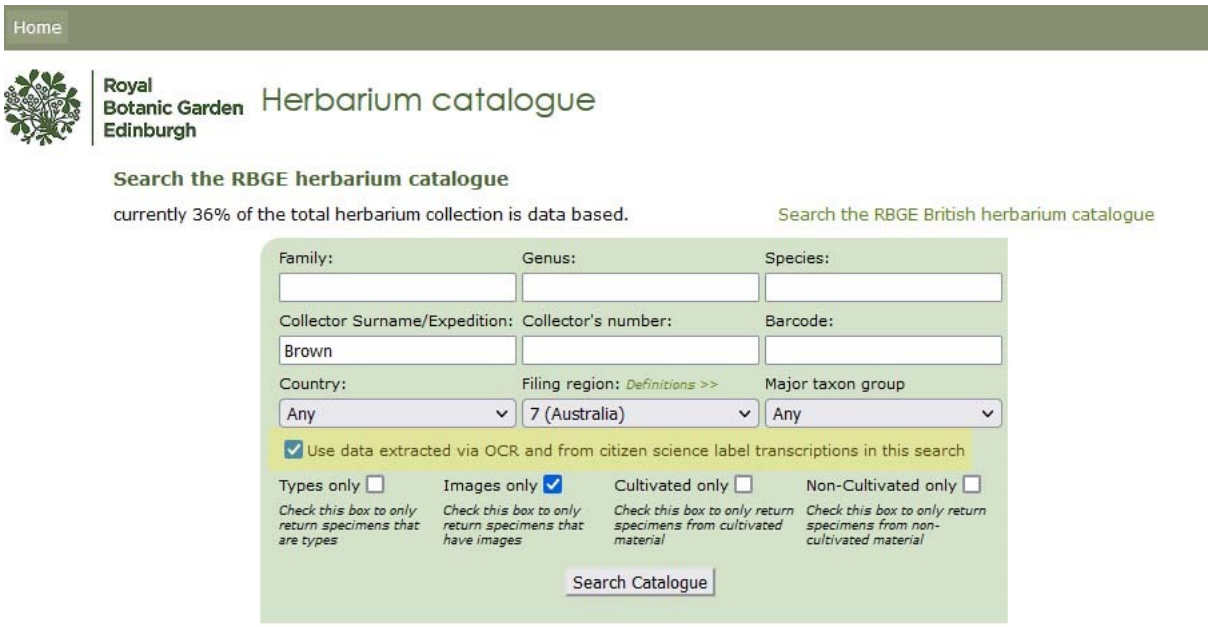

Figure 5: On the RBGE online herbarium catalogue landing page, the search term 'Brown' has been used in the Collector Surname/Expedition field and the citizen science label transcription search parameter is checked. Copyright: Royal Botanic Garden Edinburgh, 2022.[6]

[6] Royal Botanic Garden Edinburgh, *Herbarium Catalogue Landing Page,* screenshot, Royal Botanic Garden Edinburgh, 2022, <https://data.rbge.org.uk/herb> [accessed 23 February 2022].

Figure 6: A search result for specimen E00788971 on the RBGE online herbarium catalogue using the search term 'Brown' in the Collector Surname/Expedition field and the citizen science label transcription search parameter is checked. Copyright: Royal Botanic Garden Edinburgh, 2022.7

Figure 7: The Digivol (a collaboration between Australian Museum and the Atlas of Living Australia, a CSIRO hosted NCRIS facility) transcription for specimen E00788971 on the RBGE online herbarium catalogue. Any search parameters found in the transcription data are highlighted yellow, in this instance the collector surname 'Brown' is highlighted. Copyright: Royal Botanic Garden Edinburgh, 2022.8

Data quality concerns can be alleviated by providing clear guidance on what the purpose of the data collection is and what success looks like. Guidance should also support volunteers to approach any idiosyncrasies of the collection. Standardised data entry can be promoted by the use of authority and terminology controls. These analogous controlled vocabulary systems used in library (authority controls) and museum catalogues (terminology controls) already support discoverability of objects through collocation of associated items by implementing standardised form of a term, for example, the spelling of a person's name, place-name, subject descriptor or object type. A record will exist that contains the required format and associated

[7] Royal Botanic Garden Edinburgh, *Herbarium Catalogue Search Result for Brown of specimen e0078891,* screenshot, Royal Botanic Garden Edinburgh, 2022, <https://data.rbge.org.uk/herb> [accessed 23 February 2022].

[8] Royal Botanic Garden Edinburgh, *Digivol transcription of specimen e0078891,* screenshot, Royal Botanic Garden Edinburgh, 2022, <https://data.rbge.org.uk/herb> [accessed 23 February 2022].

variants.[9] The association between a standard term and its variants expands the potential of discovering and grouping associated items by making the search parameters more flexible. Indeed volunteers could actively participate in the gathering of variants to assist in the development of these controls. In addition the field of artificial intelligence in the form of machine learning is now able to help with both data annotation and validation. However its use is still limited by the requirements for large training sets. Crucially, when considering data quality it should be noted that baseline data is often variable and some organisations are often (ironically) comfortable publishing poor quality OCR data. Concerns therefore could be circumvented by providing data attribution to all sources.

Key Challenges

Misalignment between data structure of citizen research platforms and institutional CMS.
Quality of crowdsourced data often seen as a barrier to ingestion.
Lack of resources and skills to clean and aggregate data to organisationally required standards.
Multiple stakeholders with differing needs impacts post-project data flow.

Key Recommendations

Project design to include: project purpose, data ingestion pipeline and final destination.
Staff responsibilities for each stage of project to be clearly defined and agreed in design process.
Collaboration between organisations and CMS providers to develop effective data flow between systems.
Provision of interim data-sharing solutions where ingestion is not possible e.g. Github repository or adjacent database hosted alongside CMS records on the institutional discovery platform.

Data Rediscovery and Research

How do we maximise possibilities for the reuse of citizen research (or crowdsourced) data?

Challenges

There were three main areas of concern: licensing, institutional resources and discovery platform limitations.

[9] Collections Trust, *Collections Trust*, <https://collectionstrust.org.uk/terminologies/> [accessed 16 February 2022]; International Federation of Library Associations and Institutions, *International Federation of Library Associations and Institutions*, <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9031> [accessed 16 February 2022].

Several institutions have concerns over the licensing and legalities of reusing data, which in some instances are tied to income generation. The complexity of this issue increases when there are project partnerships which can create additional layers of restrictions on access.

With regards to organisational resources there can be a large imbalance between the rate of data collection by volunteers and the staff resources available to catalogue it afterwards.

It is impossible for an organisation to anticipate and design a discovery platform for every use case and thus it follows that these systems often only answer a limited range of user requirements. Data hosted on these platforms is often static, uploaded to a platform once, rather than 'live'.

Recommendations

Data usage rights should be considered when designing the project, with project teams advocating for open access to data created by citizen researchers. Where this is not possible, organisations should consider whether the use of crowdsourcing is ethical. If crowdsourcing is used in conjunction with subsequent data usage restrictions it is critical that this is clearly communicated to all stakeholders and particularly to volunteers who can then make an informed decision about participation. Wherever possible, data should be open access. Organisations that have opened up their collections data have had a very positive experience.

Reusability can be further improved by implementing standards such as authority and terminology controls providing a wider pool of search terms and therefore accessibility of the data to a broader audience. Providing multiple output formats that are compatible with different systems and use cases would also broaden access and reuse, as exemplified in the *Mutual Muses* project. Additionally 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. Providing attribution to datasets acknowledges volunteer contributions and allows its use to be traceable. The *Mutual Muses* project associated usernames with transcription outputs as a way of providing data attribution. Data visibility and findability can also be aided by the attribution of DOIs to each data set.

Continuity of relationships and engagement with volunteers beyond an individual project basis may provide solutions to some of the access and use case challenges. This could take the form of a co-created folksonomy such as a collaborative tagging system (e.g. hashtags) developed between citizen researchers and the organisation. This should then be incorporated into the institution's discovery platform, or intermediate data sharing platform where direction ingestion is not possible. This dynamic tagging system where multiple tags can be assigned to a particular item allows citizen researchers to define their own search terminologies, expanding search and discovery possibilities, and making the data useful to a broader constituency. Provided with the raw data, volunteers could create their own accessible systems for their research use cases. This could take the form of collation of project data on open knowledge-

base platforms like Wikidata and Wikimedia Commons. These platforms make data discoverable and accessible in other languages, allow mapping between different systems of identifiers (such as ISNI), and can host images and media created through data collection projects. Further still, volunteer contributions could be sought in the analysis of the transcription data. However, it must be recognised that these options would be difficult for organisations with fewer resources.

Key Challenges

Licensing and legalities of reusing data.

Data use is often tied to income generation.

Project partnerships can add additional access restrictions.

Imbalance between rate of data production and ingestion or cataloguing.

Discovery platforms only answer a limited range of user requirements.

Uploaded data is often static rather than in a live and editable state.

Key Recommendations

Data usage rights to be considered within the project design.

Wherever possible data should be made open access.

Data rights to be clearly communicated to all stakeholders at the outset.

Implement data standards via the use of authority and terminology controls.

Provide output formats that are compatible with multiple systems and use cases.

Provide training in data access and use.

Provide data attribution to acknowledge volunteer contribution and publish provenance.

Apply Digital Object Identifiers (DOIs) to datasets to aid visibility.

Engage volunteers beyond data creation by collaborating on access and use case challenges.

Engagement

How do we make crowdsourced data accessible to all?

Challenges

Audience demographics can be narrow for a variety of reasons, including: language barriers due to the anglocentric bias of crowdsourcing platforms, low confidence levels or lack of trust in technology, and lack of accessibility features for users with visual, aural or learning

disabilities. This narrow window of engagement can inadvertently privilege one kind of engagement and devalue others.

Ethical concerns were raised in terms of reaching out to underrepresented demographics within volunteer cohorts. There are instances of historic injustices carried out against these demographics that can make asking for unpaid labour inappropriate. The potential for volunteering to replace paid positions within organisations was raised.

Recommendations

Avoid delimited roles of contributor and consumer at all stages of engagement as this removes the issue of imbalance between who benefits the most from the interaction. Use a variety of media opportunities to diversify who engages with the data such as institutional membership, organisational volunteer communities, social media, sites where crowdsourcing initiatives are promoted and more traditional media such as newspapers, radio and television, which may be more widely trusted.

Crowdsourced data should be open access and provided in more than one format in order to make it accessible to a variety of audiences. Supporting documentation should provide information on the background of the project, guidance on data reuse as well as indicating the need for project acknowledgement in reuse of the data.

Key Challenges

Language barriers due to anglocentric platforms.

Low confidence levels or trust in technology.

Lack of accessibility features for users with visual or learning disabilities.

Narrow engagement can inadvertently privilege one kind of engagement while devaluing others.

Appropriateness of asking for free labour from underrepresented demographics.

Volunteering replacing paid roles.

Key Recommendations

Avoid delimited roles of contributor and consumer at all project stages to remove imbalance of power and project benefits.

Use a variety of media opportunities to diversify engagement including older more established media like newspapers, radio and television.

Data should be open access and provided in multiple formats.

Provide supporting documentation on project background, data reuse and project acknowledgement requirements.

Conclusion

Careful project design is key and critically must take into account any potential barriers to data access, use and reuse. These need to be addressed in the design stage with any restrictions clearly communicated to all stakeholders from the outset. The priority should be access before CMS integration, therefore raw data and interim repositories are acceptable stepping stones. Ideally projects should aim for open access data in a variety of formats to suit the needs of different audiences. Responsibilities for each stage of data creation and movement should be clearly identified. A more collaborative approach between volunteers and staff should be explored. Data quality concerns can be alleviated by the application of attribution and provenance to all data sets, be they produced by volunteers, staff or a machine-learning model.

Acknowledgements

Thank you to all the participants and presenters who took part in this workshop.

References

Arts and Humanities Research Council, *Towards a National Collection,* <https://www.nationalcollection.org.uk/> [accessed 22 February 2022].

Bernard Ogden, *Toward a National Collection — Engaging Crowds*, 2019 <https://tanc-ahrc.github.io/EngagingCrowds/index.html> [accessed 22 February 2022].

Collections Trust, *Collections Trust* <https://collectionstrust.org.uk/terminologies/> [accessed 16 February 2022].

Fiske, Hugh via the crowd. *Bronze Age Spreadhead SL. 746.* screenshot, Sketchfab, 2021 <https://sketchfab.com/3d-models/bronze-age-spearhead-sl746-2e87e13023774a27ac8f59c9a231cfcf > [accessed 23 March 2021].

International Federation of Library Associations and Institutions, *International Federation of Library Associations and Institutions* <https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9031> [accessed 16 February 2022].

National Endowment for the Humanities, *National Endowment for the Humanities* <https://www.neh.gov/> [Accessed 23 March 2022].

Witten, Ian H., David Bainbridge and David M. Nichols, *How to Build a Digital Library*, 2nd edn (Burlington: Morgan Kaufmann, 2010).

Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, Amy Boyer, Andrea Simenstad, *et al*, 'Individual vs. Collaborative Methods of Crowdsourced Transcription' *Journal of Data Mining and Digital Humanities* (2019) <https://hal.archives-ouvertes.fr/hal-02280013v2> [Accessed 16 March 2021].

Sciafabric, *Pybossa,* 2022, <https://pybossa.com/> [accessed 22 February 2022].

The University of Cambridge, University College London, University of Stirling, *Micropasts* <https://crowdsourced.micropasts.org/> [accessed 22 February 2022].

The Zooniverse, A screenshot of the ALICE interface with an annotated image of the letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864, and corresponding transcriptions of each line of the letter, screenshot, The Zooniverse <https://alice.zooniverse.org/> [accessed 23 March 2021].

The Zooniverse, A screenshot of the ALICE interface showing an image of a letter from Lydia Maria Child, Wyland, to Robert Folger Wallcut, Aug. 26th, 1864 and three transcriptions carried out by three different volunteers [usernames have been redacted] and an aggregate of the three, screenshot, The Zooniverse <https://alice.zooniverse.org/> [accessed 23 March 2021].

The Zooniverse, *The Zooniverse* <https://www.zooniverse.org/> [accessed 22 February 2022].

Appendix 1 Workshop participant poll

We ran a poll during the workshop to gain a better understanding of participants' experience of working with citizen research projects.