Challenges for integrating volunteer and AI-enriched metadata into collections systems

What are we studying?

There's a long history of crowdsourcing projects at cultural heritage institutions (GLAMS, or galleries, libraries, archives and museums; including citizen science and citizen history projects). More recently, GLAMs have experimented with machine learning or 'AI' to create, enrich or enhance data about collections. However, some projects struggle to integrate the data created or enriched by online volunteers and/or machine learning into collections management and discovery systems (catalogues, for short).

We're seeking to understand the barriers and successes for projects incorporating enriched data into catalogues or other core systems by gathering information on the types of data, tools and processes used by project teams. We hope these results will help organisations, software suppliers and projects with the work of integrating enriched data appropriately into collections systems.

This survey should take about 15 minutes to complete.

Who can take this survey?

We're interested in the experiences of anyone who's worked on crowdsourcing or machine learning projects to enrich collections data. We're particularly interested in hearing from projects in Europe, Asia and Africa.

This survey is designed so that more than one person can respond for any institution or project, especially for large or complex organisations. We also welcome responses from inactive projects, and past project teams. Please feel free to collaborate with colleagues, or provide individual responses. **If you can't provide a comprehensive answer for a question, feel free to provide a partial response from your own perspective.** If you have more than one significant project, **you may wish to do the survey once for each project.**

The survey is particularly designed for people working in collecting institutions (libraries, archives, museums, etc) with their own catalogues, but we also welcome responses from projects that create or enrich data through e.g. research or community projects working with data from GLAMs, or 'roundtripping' records to return enhanced data to a catalogue.

What will we do with the results, and where can you access them?
We will share the results of this research on the Collective Wisdom website (https://collectivewisdomproject.org.uk), in blog posts by the British Library and Zooniverse, and in conference papers or journal publications. If you give us your email address and permission to email you, we'll send you the results by email.

The survey has been designed so that you can answer anonymously if you wish and will not be identifiable by those with access to the data. There is also an opportunity to provide contact data if you wish, but we will not publish this information.

Who are we? And who can you contact with questions?

You can email us (Mia Ridge, Sam Blickhan and Meghan Ferriter) via digitalresearch@bl.uk. We were the investigators on the AHRC-funded Collective Wisdom project. After launching our white paper 'Recommendations, Challenges and Opportunities for the Future of Crowdsourcing in Cultural Heritage: a White Paper' last year we've re-convened to explore this question, as it's so fundamental to ensuring the benefits of crowdsourcing and AI work.

The survey is open until April 18th, but we encourage you to complete it sooner!

About your project

1. Project name

   _____

2. Project URL

   Even if it's no longer active

   _____

3. Project start date, end date

   If applicable

   _____

4. Which is the best description of your project? It produced or enriched collections data through...

   Mark only one oval.

   - ◯ crowdsourcing (without machine learning / AI in the same project)
   - ◯ machine learning / AI (without crowdsourcing in the same project)
   - ◯ a combination of crowdsourcing and machine learning / AI
   - ◯ Other: _____

5. Who manages the definitive version of the collections records you're working with?

   Mark only one oval.

   - ◯ Our department / team
   - ◯ Another department / team in our organisation
   - ◯ Another organisation
   - ◯ Other: _____

6. Which crowdsourcing or machine learning platforms or tools does your project use?

   _____

7. What kinds of data does the project create or enhance?

   For example, whole or selective transcription, tags or captions, classifications or labels, translations, etc.

   _____

   _____

   _____

   _____

   _____

8. How was data created or enhanced?

   For example, crowdsourced or machine learning task transcription, tagging, etc

   _____

Has your project been able to integrate enriched data into collections systems?

We're particularly interested in the technical and organisational factors that support the import of new metadata and/or updating existing records

9. Have you been able to integrate enriched data into catalogues?

Select 'Other' for partial ingest, or projects where ingest / integration wasn't the goal. Here 'enriched data' means records created or updated outside a core collections system

Mark only one oval.

- ◯ Yes - but new records only
- ◯ Yes - but updated records only
- ◯ Yes - new and updated records
- ◯ No
- ◯ Other: _____

10. Which catalogue application are you working with?

That is, what's the name of your internal or open source software, commercial catalogue? Optionally, who supplies it? e.g. vendor name.

_____

11. In your experience, does anything about your catalogue software support or hinder the ingest of enriched data?

_____

12. If you were able to integrate some or all enriched data, briefly describe the process

_____

_____

_____

_____

_____

13. Were any factors particularly important for enabling data ingest?
    E.g. data export formats, data standards, APIs, agreements between departments

14. If you weren't able to integrate some or all enriched data, why not?

15. If you're still in the middle of an ingest process, what have you learned so far? What is still a challenge?

16. Does updating existing records present different challenges than importing new ones?

17. Do you use any 'middleware' systems to store and/or display data?

For example, systems that store additional information or links between records in separate applications, e.g. collections records and digital asset management, or a shop / licensing system and collections records

_____

Quality assurance for enriched data

Do you check the quality of crowdsourced or machine learning enhanced records in some way?

18. Does your project have any automated or manual processes for checking enriched data before, during or or after ingest?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

19. If so, how do you check the quality of enriched data?

_____

20. What does 'data quality' mean for your project?

E.g a project-, institution or sector-specific definition or standard

_____

21. If you use machine learning or 'AI', do corrections to records help improve the model?

E.g. human in the loop, reinforcement learning from human feedback

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Not applicable
- ◯ Other: _____

22. Can users or the public report incorrect data and/or suggest improvements?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

Recording the provenance of enriched data

If you're able to ingest data, is some record of its source kept?

23. If enriched data is incorporated into existing records, can **staff** tell the difference between any original data and the enrichments?

_____

_____

_____

_____

_____

24. If enriched data is incorporated into existing records, can **the public** tell the difference between any original data and the enrichments?

About your project / organisation

We're particularly interested to know how many different individuals and teams are involved in a single ingest / data project. Please feel free to forward this survey for a wider range of responses.

25. How many people were involved in your project, and what were their roles?

26. What kind of institution is your project based in?

For 'Other' - this might include a university, community organisation, etc. If you've already filled this in once for a different project, you can skip it.

Mark only one oval.

- ◯ Library
- ◯ Archive
- ◯ Museum or art gallery
- ◯ Other: _____

27. Which city or region is your project / organisation based in?

28. Which country is your project / organisation based in?

29. How many paid employees does your organisation have?

A rough guess is ok, it's just to give us a basis for comparison

Mark only one oval.

- 1-4
- 5-19
- 20-49
- 50-99
- 100-249
- 250 to 499
- 500 or more

About you

All questions in this section are optional. Information provided here will be used to reach out with follow up questions and/or notification of results only with permission. This information will not be included in published results.

30. Your name

31. What department or area of your organisation do you work in?

32. Your email (if you'd like the results and/or are ok with follow-up questions)

33. Can we contact you by email to share our survey results?

Mark only one oval.

- ◯ Yes
- ◯ No

34. Can we contact you by email to ask follow up questions?

Mark only one oval.

- ◯ Yes
- ◯ No

Is there anything else you want to tell us?

35. Is there anything else you want to tell us about your project or institutional approaches to crowdsourcing/AI/machine learning?

E.g. policies, challenges, examples you want to share. You can also email us via digitalresearch@bl.uk