ONLINE SUBMISSION TO NESTA IN RESPONSE TO CALL FOR EXPRESSIONS OF INTEREST: COLLECTIVE INTELLIGENCE GRANTS EOI STAGE 1A
SUBMITTER: CHRONOSCOPIC EDUCATION
DATE: FRIDAY, NOVEMBER 11TH, 2018

Background to call: https://www.nesta.org.uk/project/collective-intelligence-grants/call-ideas/

YOUR CONTACT DETAILS

1. Lead Organisation name:

Chronoscopic Education

2. Lead Organisation Type:

Other - write in (Required): Charitable incorporated organisation [Board formed & registration process underway]

3. Registered charity number (if applicable)

N/A

4. Registered company nuimber (if applicable)

N/A

5. Registered address of lead organisation

16 The Avenue
London
N10 2QL
United KIngdom

6. Website of lead organisation:

http://Chronoscopic.org

7. Lead organisation contact person :

Mr Colin Greenstreet

8. Lead person email:

colin.greenstreet@gmail.com

9. Lead person phone number:

0044-(0)7340-119492

PROPOSAL

10. Proposal name:

ArchiveBot

11. Proposal summary (please provide a summary of your proposal, no longer than 150 words):

We propose an experiment to build an ArchiveBot and related collective intelligence processes for the semi-automatic extraction of metadata from handwritten manuscript material, using C17th Admiralty Court depositions held at UK National Archives as the use case. Experiment is supported by UK National Archives.

Our experiment will involve facilitated teams of archivists, volunteers and machine learners working to create useful metadata. It is a POC of an approach we will build on in 2020 and beyond in the Signs of Literacy initiative.

Specifically, we will use machine learning and volunteers, working iteratively, to identify text areas containing "front matter" (names, occupations, age...).
[Example: http://www.marinelives.org/wiki/HCA_13/68_f.1r_Annotate]. We will fit boundary boxes (manually/automatically), working with facilitated teams, and will "read" the boxes semantically, using humans, named entity recognition and HTR. We will build on prior work with NER, HTR and Key Word Spotting.

Background:
https://github.com/Signsofliteracy/ArchiveBots/blob/master/ArchiveBots_Discussion_Document_Ver3.9_13102018.pdf

12. Please explain what you are wanting to test/learn through your experiment(s)? And what outcome measure(s) do you envisage using to answer this? (max 150 words) :

POC of iterative, collaborative working of archivists, volunteers and machine learners. Learning about process will be as important as learning about, and development of, tools. The training, skills and culture of these three groups are very different. Our experiment will create a structured way for individuals to work closely together, both face to face and long distance (US, UK, Europe), and will enable us to better structure the whole Signs of Literacy initiative as well as the specific metadata extraction workstream of which the ArchiveBot experiment is part.

Test: (1) How best to structure machine and human workflows? (2) Most effective ways of presenting machine generated data to archival and non-archival volunteers for human processing and return to machine learners for further work? (3) Sufficiency of available technical tools? (4) Incentivisation and satisfaction?

Outcome measures: (1) Participant qualitative survey (2) Quality & quality of output (3) Productivity

13. How do you plan to carry out your experiment? (max 250 words)

We have an outline experimental process and design, which we are continuing to develop, and we hope with input from NESTA.

Input received from John Sheridan, digital director of the TNA; Catharina Gronqvist, AI coordinator, Swedish National Archives; Dr Walter Reade, Kaggle data scientist; Glen Robson, technical

coordinator, IIIF consortium; Dr Ben Albritton, Stanford Libraries; Dr Mark Hailwood, University of Bristol. We have also formed two volunteer panels (archival; machine learning) to support the broader initiative. Several members of these panels have worked previously as volunteer transcribers and annotators for the MarineLives and other Chronoscopic Education projects.

Outline plan

(1) Run Kaggle Signs of Literacy competition (March - April 2019), on which the ArchiveBot project will build.

(2) In parallel, convene the ArchiveBot Steering Committee (April 2019) to (a) agree experimental goals, output measures, full time and part-time team membership and correct balance of machine learning, archival and volunteer members. The Steering Committee will meet virtually every six weeks.

(3) Form team, which will be physically centred on the National Archives UK (archival and technical), but will draw on IT and machine learning expertise from elsewhere in UK, Europe & US.

(4) Recruit and train archivist and general public volunteers for collective intelligence teams.

(5) IIIF infrastructure & development resource to be provided by Dutch digital firm Picturae (subject to discussion of terms).

(6) Team to be project managed by Colin Greenstreet (Chronoscopic Education), who will dedicate approximately three days/week to the project.

(7) Run second Kaggle competition (?November/December 2019) testing ArchiveBot technical solutions.

14. Please outline how this will help create actionable insights for practitioners and have wider applicability? (max 150 words)

Large scale digitisation of handwritten historical manuscripts has taken place over the last ten years in archives worldwide. Specific digital series are as large as one to ten million images. Most of images lack meaningful metadata to enable search by researchers, archivists and general public, and are almost entirely untranscribed.

Machine learning techniques offer an opportunity to improve searchability of untranscribed images, and to enable the intelligent synthesis of information from these images.

Full text HTR is one machine learning approach, which is proving fruitful [READ/Transkribus; other initiatives; in which we are active]. However, intelligent synthesis of information from multiple parts of a digitised document, is a holy grail, which has not been meaningfully addressed.

POC will provide a template for (1) Structuring workflows (2) Presenting data (3) Human/Machine dynamics. The insights will impact the design and execution of our broader initiative, and will be shared widely by us.

15. Please describe if your experiment is part of a larger practical project or research programme (max 150 words)

ArchiveBot is an important early collective intelligence experiment in the Signs of Literacy initiative's metadata extraction and process development workstream.

Our initiative will run from 2019 to 2023 and has five objectives. (1) Development of tools and processes for increasingly automated metadata extraction and linkage from digitised handwritten manuscript pages (2) Development of infrastructure to deliver a highly scalable data set of one million manuscript pages containing "signoffs" (marks, initials, signatures) and related front matter (3) Stimulating the development of machine learning & collective intelligence capability in archival centres of excellence in UK and elsewhere (4) Building interest amongst historians and linguists in machine learning applications to support close and distant reading approaches to large scale data sets

Our six workstreams are (1) Financing (2) IT infrastructure & standards (3) Content and technology partner development (4) Content identification and acquisition (5) Metadata extraction tool and process development (6) Kaggle competitions.

16. Do you anticipate requiring support from Nesta to help with the design and/or delivery of your experiment?

Yes

YOUR TEAM

17. Please give details of any partner organisation involved in this proposal (if applicable)

Name of Organisation: The National Archives, UK [TNA digital director involved in shaping proposal, but formal participation in a bid will require a six week approval process]

Type of organisation: National archive; lead responsibility for UK archival sector; driving AI in UK archival sector

Organisation website: http://www.nationalarchives.gov.uk/

18. Please tell us briefly about your team and its experience relevant to your proposal (max 150 words)

Colin Greenstreet. Signs of Literacy community organiser; trustee of Chronoscopic Education; founding director of Marine Lives. [https://www.linkedin.com/in/colin-greenstreet-7434b9/]

* Extensive experience of managing complex, technically demanding projects (Booz.Allen & Hamilton; R&D strategy director & Vice President at GlaxoSmithKline). Likewise, originating & running collaborative, international digital projects bringing together volunteers, academics, archivists and technologists using innovative facilitated teams for expert/crowdsourcing [http://www.marinelives.org/wiki/MarineLives; https://github.com/Signsofliteracy/Signoff]

Dr Mark Hailwood. Lecturer, social history, University of Bristol. Academic interest in historical literacy. Co-founder of Signs of Literacy in March 2018. Engaged with Colin in shaping initiative. Identifying key archival collections; forming cross-disciplinary academic networks to support

initiative; conceptualising ways to measure sophistication of marks, initials and signatures; driving development of content research projects to use growing dataset.

John Sheridan. Digital director at the National Archives, UK (TNA). Steering committee membership. Confirmed. [http://www.nationalarchives.gov.uk/about/our-role/executive-team/john-sheridan/]. Additional technical team member(s) to be supplied by TNA. Considering adding UK machine learning academic to team.

Catharina Gronqvist. AI coordinator, Riksarkivet (Swedish National Archives). Archivist; digital science specialist. Steering committee membership. Confirmed.

Kaggle data scientist representative to be selected in discussion with our partner Kaggle. TBC. Kaggle, the world's largest machine learning community, is providing pro bono competition and data science services to the Signs of Literacy initiative.

IIIF representative to be selected in discussion with IIIF consortium. TBC. Probably Glen Robson, Technical Coordinator, IIIF consortium. Strong systems, crowdsourcing, and archival experience, with a Masters in Software Engineering, experience in repository development, digital asset management, and as head of systems at National Library of Wales.

BUDGET

19. What is the total funding that you are requesting from the collective intelligence experimentation fund? (the limit is £20,000, but you can ask for anything up to that amount)

£20,000

20. Do you need any additional funding to run your experiment(s)

Yes, and I am currently looking for additional funding

KEEPING IN TOUCH

We are creating a community of people who are interested in collective intelligence. We would like to stay in touch to share updates from the Centre for Collective Intelligence Design and Nesta's other work, including regular newsletters, jobs, funding opportunities, programme updates, new research and publications, invitations to events and the occasional request to take part in research or surveys.

You can unsubscribe by clicking the link in our emails, or emailing info@nesta.org.uk. We promise to keep your details safe and secure. We won't share your details outside of Nesta without your permission. Find out more about how we use personal information in our Privacy Policy.

Your answers in this section will not affect your application.

21. I am happy to be contacted with updates from Nesta's Centre for Collective Intelligence Design

YES

22. I would like to hear more about Nesta's other work?

YES

23. Please tell us which other areas of work you are interested in

Education

USE OF PERSONAL INFORMATION

This grant programme is being run by Nesta. Your personal information will be used by Nesta to review your application and for ongoing management and administration of the grant programme. Your details and information will be shared with Nesta personnel who are managing the administration of the grant programme. You can find details of the Nesta team here. Nesta may also share your details and information with an applicant review panel who will help us assess applications. Details of the members of the panel will be notified to you here when known. Our legal basis for doing this is our legitimate interest of being able to work collaboratively with other organisations to deliver this grant programme.

We use a third party provider, SurveyGizmo, to provide this Expression of Interest form. SurveyGizmo operates in the U.S. and complies with the EU-U.S Privacy Shield Framework. Please see their privacy notice for further information.

If you wish to know more about how Nesta will use your personal information please see our privacy policy.