Zooniverse is the world-leading platform for people-powered research, with over 50 projects across the disciplines reaching over 1.5 million volunteers and producing data that has been used in more than a hundred peer reviewed papers. While crowdsourcing projects have proliferated in the last decade, few offer easy to use, open source, and free platforms on which galleries, libraries, archives, and museums (GLAM), academics, and amateur users can rely. Zooniverse has the infrastructure, community, and technical expertise to intervene at this critical stage.

In this three-year IMLS supported effort, we propose to:

- Create 4 bespoke GLAM Zooniverse projects with improved full text transcription and audio transcription crowdsourcing tools that optimize volunteer engagement and retention while producing research-supporting output
- Implement a GLAM-friendly data pipeline for automating ingestion of digital collections into Zooniverse and facilitating the ingestion of Zooniverse outputs into GLAM Content Management Systems
- Identify successes in existing GLAM projects, and these new bespoke projects, and build these into our DIY Project Builder, so that anyone can use these tools for developing their own crowdsourcing project
- Create a small number of design templates research teams using the Project Builder will be able to choose from, depending on the project type
- Communicate research results and train the GLAM community in using the Project Builder

Through this effort, we will explore the following research questions:

- How can crowdsourcing be deployed in the arenas of text and audio transcription and metadata extraction for the purposes of unlocking big data currently trapped in GLAM sources that cannot be machine read? What methods produce the best data and make for the best user experiences?
- Does the current Zooniverse methodology of multiple independent transcribers and aggregation render better results than allowing volunteers to see previous transcriptions or markings by other users, and aggregating these results? How does each methodology impact the quality and depth of analysis and participation?
- How can we extend our crowdsourcing expertise to more GLAM professionals and learn from them, in turn, how to adjust the Zooniverse platform to best meet their research needs?

The Zooniverse is supported by software developers, educators, researchers and program managers at the Adler Planetarium and the University of Oxford. We will harness our international collaboration and expertise in order to focus on the needs of the U.S. GLAM community, but in the knowledge that the fruits of this project will support GLAM, school and research communities all around the world. We will work closely with our four bespoke project partners and advisory board throughout the project. In order to obtain input and feedback from the broader GLAM community, we will host feedback sessions at conferences to help guide us in determining which tools/approaches in existing and new bespoke projects proved successful and merit building into the Project Builder. In Year 3 we will host Project Builder training workshops at GLAM conferences to support GLAM professionals in building their own crowdsourcing projects using these tools, as well as present the results of our research efforts.

Proposal Narrative

1. STATEMENT OF NEED

As libraries, museums, and other cultural repositories digitize their collections and place them online, the challenges of transforming these materials into useful and searchable sources of information are becoming increasingly apparent. While OCR and handwriting recognition technology have opened up some print and manuscript corpora, and image and voice recognition software are improving daily, there are still many tasks that require human intervention. For these, volunteer crowdsourcing is a viable and vibrant solution.

While crowdsourcing projects have proliferated in the last decade, few offer easy to use, open source, and free platforms on which galleries, libraries, archives, and museums (GLAM), academics, and amateur users can rely. Zooniverse has the infrastructure, community, and technical expertise to intervene at this critical stage. Founded in 2007 with a single crowdsourcing project called Galaxy Zoo, Zooniverse has since partnered with academics and GLAM professionals around the world to build fifty crowdsourcing projects, resulting in over 100 publications. The projects cover a range in GLAM and Humanities subjects (from papyrology to art history to renaissance English literature) and STEM subjects (from astrophysics to ecology to climatology). Zooniverse is unique among the open source, free, crowdsourcing options as a result of its 1) shared software, experience, expertise, and input from users across the disciplines, 2) reliable, flexible, and scalable back-end, which can be used for a variety of development tasks, 3) free, do-it-yourself (DIY) 'Project Builder' (also known as Panoptes) capabilities as described below, and 4) the scale of its existing audience of participating volunteers, drawing on 1.4 million volunteers worldwide. The impact of sheer numbers is impressive; for example, the recently launched Shakespeare's World[1] Zooniverse project attracted in its first eight days 1,440 volunteers who transcribed 1,495 pages.

Zooniverse projects are not only designed to facilitate research through volunteer effort, but to create opportunities and inspiration for learning, foster co-creation and publication by members of the public, researchers and GLAM professionals, and further research into crowdsourcing as a methodology and social phenomenon. Each project has an active discussion forum called 'Talk' where volunteers and researchers can exchange ideas and ask questions. A Zooniverse-wide Talk facilitates exchange across projects. The Talks are where numerous discoveries have been made, ranging from new word variants for the *Oxford English Dictionary* to odd galaxy types. These community-focused features, core to Zooniverse, speak directly to the themes and challenges of the National Digital Platform, including the call for technology that supports "lifelong learning, and cultural and civic engagement". In contrast with this volunteer-driven, collaborative environment, a number of crowdsourcing platforms rely on radically different models. Amazon's Mechanical Turk, for example, pays volunteers to carry out microtasks, an unsustainable model for GLAM organizations.

Building on eight years of experience, in July 2015 Zooniverse launched a new version of its platform supporting a free, DIY Project Builder. The Project Builder is transformative; prior to its development a typical project required professional web development and even rapid projects took months to develop. Now anyone can build and deploy a project in an hour, using a set of browser-based tools. The Project Builder supports the most common types of interaction including classification, multiple-choice questions, comparison tasks, marking and drawing tools, or any combination thereof. Since its launch, over 2000 DIY projects have been created and 150 have generated classifications (note this number includes both public and private projects; e.g., projects in classrooms or internal to a research team). For example, Digging the Potomac[2] uses the line tool to gather measurements of pottery fragments and multiple choice questions to gather data about color, pattern, and type. Using data generated by volunteers in just one week, the project lead disproved his hypothesis that Native American pottery styles remained highly localized in the 17th century. Crowdcrafting.org similarly provides free online tools for building crowdsourcing projects, but their community is much smaller, their support for research teams less advanced, and they lack a consistent approach to aggregating or validating data.

[1] http://www.shakespearesworld.org/
[2] https://www.zooniverse.org/projects/sportprof/digging-the-potomac#/projects/sportprof/digging-the-potomac/

One of the chief concerns of academics and GLAM organizations is that crowdsourcing might generate incomplete or inaccurate data. In systems where a given task is only performed by one person, this may be a genuine problem. Instead, the Zooniverse approach rooted in scientific analysis mitigates against a 'bad' or inaccurate classifier by requiring multiple people to independently perform the same task on the same object. At least three people transcribe each page in Shakespeare's World and AnnoTate, for example. If three people transcribe the same line, these are greyed out, while outstanding lines are available to future site visitors. We then aggregate the multiple responses together using in-house algorithms. To date, the average rate of agreement among users for a line (from manuscript pages) transcribed by multiple people on AnnoTate is ~95%. This rate of accuracy is in line with estimates on OCR accuracy, which are ~94% for grayscale and ~97% for bitonal[3].

This approach is radically different from that used by other crowdsourcing platforms. For example, FromThePage and Scripto crowdsourced projects are set up so that volunteers can edit as well as add to previous volunteers' efforts. The results are not aggregated. Rather, the most recent edit is shown and previous versions are accessible. There are no data-driven algorithms for aggregation or deriving confidence metrics for consensus. These transcription platforms also do not provide a free Project Builder interface for those without technical software expertise to set up their own crowdsourcing project. We note that there are a number of important lessons learned from these projects which do inform our efforts, and highlight them in Section 3.

Here we propose a research and development program to identify and implement crowdsourcing best practices in the arenas of text and audio transcription for the purposes of unlocking big data currently trapped in GLAM sources that cannot be machine read. The Project Builder is a flexible, powerful, and easy-to-use resource with a wide range of potential applications for GLAM collections. However, it does not yet incorporate full text transcription tools, audio transcription tools, or a GLAM-friendly data pipeline, functionalities most requested by GLAM institutions. We propose to build four bespoke projects in order to test, iterate, and research the efficacy of new and existing approaches (including within current Zooniverse and other projects) in these arenas. From these lessons learned, we will add new tools and functionality to the Project Builder. Zooniverse has successfully used this model of building bespoke projects in order to learn which tools are most useful, before implementing tools in the Project Builder. We recognize that volunteers' "cognitive surplus"[4] is precious, and are therefore unwilling to waste it with tools that are not proven to extract data in an efficient, high quality, and useful form.

2. IMPACT

Addressing GLAM Concerns/Issues for Long-term Impact: Through this proposed effort we will directly address the following needs of the GLAM community:

- Full text and audio transcription tools and interface designed in collaboration with GLAM institutions that optimize volunteer engagement/retention while producing research-supporting output
- Image annotation/marking tools that encourage volunteer engagement/retention while producing research-supporting output
- The foundation for a GLAM-friendly data pipeline for ingestion of digital collection materials into Zooniverse projects and facilitating the uptake of Zooniverse outputs into GLAM Collections Management Systems (CMSs) like Islandora and Omeka
- An enhanced DIY Project Builder supporting GLAM professionals in building their own crowdsourcing projects for free
- Hands-on, in-person training in using the Project Builder
- Research focused on GLAM crowdsourcing approaches, including investigating the impact of transcription with and without knowledge of others' input

[3] Powell, Tracy, and Gordon Paynter, 'Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images', *D-Lib Magazine*, Vol. 15 No. 3/4 (2009), <http://www.dlib.org/dlib/march09/powell/03powell.html>

[4] A term borrowed from Clay Shirky's rousing 2010 TED talk, in reference to the trillion hours of leisure time humans collectively have each year, which could be harnessed to advance knowledge through civic engagement.

Inviting Input, Consensus-Building, and Buy-in: We will work closely with four different GLAM institutions to build four bespoke crowdsourcing projects (two projects for text transcription and two projects for audio transcription, with details on each below). One bespoke project is already identified. The remaining three will be identified through open calls. We will host feedback sessions at GLAM conferences for input from a broader set of institutions on our decisions and approach for building in new functionality into the Project Builder. In Year 3 we will host training workshops for GLAM professionals in using the Project Builder to build their own crowdsourcing projects, incorporate the results into their databases and research, and sustain and nurture their online volunteer communities.

Performance indicators, targets, metrics, and measuring success: In order to gauge impact on the accessibility, discoverability, and searchability of digital content and the improvement of our partner institutions' digital library infrastructure, we will use the following:

(1) *Metadata creation/enrichment:* Each partner institution will provide an assessment of the existing record details for their digital collection. We will then measure the number and extent of the metadata fields added, the quality of metadata fields added (e.g., word error rate), and the extent to which our partner institutions' CMSs incorporate the newly crowdsourced results.

(2) *Increased digital collections usage:* Each partner institution will provide available metrics on current levels and breadth of usage of their collection, including page hits, audience reached (both geographic and by institution; e.g., GLAM institutions, schools, individuals, etc.), citations in published works, and use in teaching curricula and other resources. We will use Google Analytics and create a survey for partner institutions to capture quantitative measures of change in these metrics, pulling from the "Toolkit for the Impact of Digitised Scholarly Resources"[5] and the "Guide to Impact and Audience Analysis"[6].

(3) *Improved digital library infrastructure:* We will create a survey for partner institutions to report on the impact on strategic initiatives within their institution as a result of participating in this effort, including in digital asset management, conservation, collection management, and public engagement.

Assessment Plan: For each bespoke project, we will create an individual assessment plan based on the information listed above, and timeline for gathering and analyzing the metrics. For Project Builder projects, we will request that institutions complete our surveys and we will use Google Analytics on our project sites to measure impact. Our assessment of the Year 1 text bespoke projects will be used to inform the Year 2 audio bespoke projects. All performance metrics will inform and guide our Project Builder tool and interface building efforts and the design of our Project Builder training workshops.

We will gather participation metrics from our feedback sessions and workshops, i.e., participant numbers and demographics (host institution, geographic location, etc.). Workshop participants will also complete surveys measuring increase in understanding, interest, and confidence in using our crowdsourcing tools.

Tangible Products include: 1) four bespoke GLAM crowdsourcing projects, built, supported, and maintained by the Zooniverse; 2) crowdsourcing tools for text transcription, audio transcription, and image annotation/marking with accompanying research publications providing metrics and reporting methodologies for new approaches and new datasets, and 3) an expanded Project Builder supporting the successful tools and approaches identified through this effort for anyone to use in building their own crowdsourcing projects.

Sustainability: The Zooniverse is a long-standing platform with an excellent track record of supporting ongoing projects. We therefore expect to be able to support the bespoke projects built under this call for their full lifetime. The incorporation of tools into the core Zooniverse platform, which is relied on by a variety of

[5] http://microsites.oii.ox.ac.uk/tidsr/welcome
[6] https://www.jisc.ac.uk/guides/measure-the-impact-of-your-digital-resources

institutions and includes the Project Builder, also contributes to the sustainability of our efforts. All code is released under an open source license, and is available for institutions to adopt should the core platform not be available. See the Digital Stewardship document for additional details.

3. PROJECT DESIGN

Project Goals and Objectives:

- Create four bespoke GLAM Zooniverse projects that explore improvements to full text transcription and audio transcription crowdsourcing tools
- Lay the foundation for a GLAM-friendly data pipeline for automating ingestion of digital collections into Zooniverse and facilitating the ingestion of Zooniverse outputs into GLAM CMSs
- Identify successes in existing GLAM projects and these new bespoke projects and build these into the Project Builder, so that anyone can use these tools for developing their own crowdsourcing project
- Research differences in impact between transcribing in isolation versus with knowledge of how others have transcribed the same document
- Communicate these results and train the GLAM community in using the Project Builder

Activities Supporting Project Goals and Objectives:

Years 1, 2, 3 - Research: Described in detail in the section below.

Years 1 and 2, Building Four Bespoke Projects: Through this proposed effort, we will build four GLAM crowdsourcing projects to explore the specific needs and best practices for text transcription and audio transcription in the context of GLAM digital collections. We will build two text transcription projects and two audio transcription projects. **In the Research section below we provide details on the bespoke project tools, functionalities, and approaches we propose to explore.**

Because Zooniverse is regularly approached by GLAM institutions (~6 per week) with viable and interesting projects, we have already identified a bespoke full text transcription project to begin on Day 1 of this proposed effort. Prior to the start of the grant period, we will announce an open call to identify the additional text bespoke project. We will build both text transcription projects in Year 1.

The open call for the two audio bespoke projects will occur at the end of Year 1, and we will begin the build in Year 2. We will thus be able to take advantage of lessons learned through the Year 1 projects and spread project management efforts (in particular, liaising with the bespoke partner institutions) over a longer time period. This also has the positive benefit of staggering the launch of the four projects. A priority for both open calls is to reach out directly to minority-serving institutions and projects addressing issues for minority audiences. Among our Letters of Support is a commitment from the Univ. of Minnesota to apply during our open call for two of their oral history collections: one on immigration and the other on the Jewish experience.

Year 2, Broader Community Input/Feedback: We will host feedback sessions at conferences to gather input from the broader community. Community, bespoke partner, and advisory board feedback will guide us in determining which tools/approaches proved successful and merit building into the Project Builder.

Year 2, Expanding the Project Builder: The development effort to build new tools/functionality into the Project Builder encompasses both building the front-end and back-end code to support these tools/workflows as well as creating a user-friendly Project Editor interface. Key to the continued success of our Project Builder is the ease-of-use and simplicity of the web-based Project Editor interface. Through this proposed effort, we will also create a small number of design templates research teams will be able to choose from, depending on the project type. This will improve the visual appeal and user experience of Project Builder built projects.

Year 3, Project Builder Training: In Year 3, we will host workshops at conferences and summer schools targeting GLAM institutions with collections for which text transcription, audio transcription, or image annotation/marking are of interest. We include image annotation/marking because those tools are already available through our Project Builder. This includes the HILT Summer School, American Library Association conference, Museums and the Web conference, the Digital Humanities Summer Institute in Canada, the Digital

Humanities conference, the Digital Humanities Oxford Summer School, the Society of American Archivists Conference, and the Midwest Archives Conference. Additionally, for image annotation/transcription, this includes the Scientific Instrument Commission conference, for which Dr. Raposo has external funding to attend each year, the Notre Dame History of Astronomy workshop, and the History of Science Society annual meeting. The non-U.S. based conferences listed here have high attendance from U.S.-based GLAM professionals. We will also host monthly Skype sessions to respond to Project Builder questions from GLAM institutions.

Research Questions and Methods: Zooniverse is poised to serve as a transformative element in the future of GLAM research efforts and the national digital platform envisioned by the IMLS. To best serve the GLAM community, we need to understand what crowdsourcing tools and data outputs would be most useful for opening up GLAM corpora. We will address the following research questions:

- Q1: How can crowdsourcing be deployed in the arenas of text and audio transcription and metadata extraction for the purposes of unlocking big data currently trapped in GLAM sources that cannot be machine read? What methods produce the best data and make for the best user experience?
- Q2: Does the current Zooniverse methodology of multiple independent transcribers and aggregation render better results than allowing volunteers to see previous transcriptions by others, and aggregating these results? How does each methodology impact the quality and depth of analysis and participation?
- Q3: How can we extend our crowdsourcing expertise to more GLAM professionals and learn from them, in turn, how to adjust the Zooniverse platform to best meet their research and curatorial needs?

Addressing Q1 (Crowdsourcing for GLAM): Only a platform like Zooniverse, supporting a diverse range of projects, can systematically address a question such as Q1. The infrastructure enables us to develop projects with differences aimed at understanding specific interventions. In the sections below, we provide details on the four bespoke projects (2 in text transcription and 2 in audio transcription) tools, functionalities, and approaches we propose to explore in order to address this question. We also describe our plans for the existing Project Builder image annotation/marking tools and for laying the foundation for a GLAM-friendly data pipeline.

Full Text Transcription: The Papers of Abraham Lincoln[7] is a digital collection of over 200,000 documents (gathered through the National Archives, Library of Congress, private collections, etc.) that will serve as the foundation for one of the two bespoke text transcription projects. The project is led by Dr. Stowell of the Abraham Lincoln Presidential Library and Museum. Dr. Stowell is an ideal project partner for this effort: he is local to Illinois, already collaborating with Zooniverse on an unrelated bespoke project ('Decoding Civil War Telegrams' https://www.zooniverse.org/projects/zooniverse/decoding-the-civil-war), is very familiar with the opportunities the Zooniverse platform enables, and is already leading a group of volunteers in text transcription for The Papers of Abraham Lincoln.

This bespoke project will provide full text transcription for the ~77,000 handwritten documents in the 'Presidential Papers' archive. These include letters, petitions, pardons, proclamations, endorsements, etc. The average document length is 2-3 pages. Currently a small group of volunteers are transcribing some of the Presidential Papers. Over the past 3 years, they have transcribed 5,500 documents, with one volunteer contributing 20% of the effort. The volunteers use an XML markup editor and output into a CMS. The quality varies across volunteers. The original goal was for an 'expert' to proofread each full transcription. However, limited staffing has meant the transcriptions are spot-checked at most. This difficulty in providing robust quality control is typical for staff-limited GLAM institutions using this approach for full text transcription. Our radically different approach avoids this roadblock. As described in Section 1, we aggregate the results from multiple volunteers independently transcribing the same line and provide confidence metrics on the consensus results. The load is thus shifted from the GLAM professionals to our large volunteer base, and publishing the confidence measures alongside the crowdsourced results provides transparency in quality across the collection.

[7] http://papersofabrahamlincoln.org/

The new interface for this project will expand on the lessons learned through current Zooniverse bespoke text transcription projects, including Ancient Lives, AnnoTate, Old Weather, Measuring the ANZACs, Shakespeare's World, Science Gossip and Operation War Diary, as well as from external text transcription projects including Transcribe Bentham, FromthePage, and Scripto. Features optimizing volunteer engagement and retention will include:

- Volunteer choice: volunteers choose which document to transcribe and can transcribe as little as a single word or as much as an entire document. We have found through AnnoTate and Shakespeare's World that micro-tasking and choice mitigates against forced or uncertain readings and retains volunteers.
- Keeping the task simple: volunteers drop points at the start and end of individual lines of text (not grammatical sentences) and transcribe between these two points. They do not use XML markup itself, which has proven to be a major repellent to participants in other text transcription crowdsourcing projects[8]. Instead, volunteers highlight words within the transcribed line and choose among different features (e.g., insertion, deletion, expansion, etc.). We will use these tagged words in each line to create the TEI markup on the back-end, for output into commonly used CMSs. While AnnoTate and Shakespeare's World record XML-like tags in their data exports, the two bespoke projects proposed here will enable us to explore best approaches for creating the TEI markup and outputting to CMSs.
- Narrowing the content focus to support sense-making: In Shakespeare's World, the first release (or 'chapter') consists of recipes and letters, with more genres to follow. Similarly, we will subdivide the 'Lincoln Papers' into 'chapters' along chronology or topics. This supports volunteer narratives connecting documents and enables 'experts' to more easily foster and contribute to discussions in Talk. We will also incorporate best practices from FromThePage in its sophisticated approach to dynamically tracking and indexing references to terms as volunteers transcribe documents.

Features optimizing use of volunteer time given the reality of limited GLAM expert staff time will include:

- Consensus through numbers: Three or more volunteers transcribe each line, after which it is no longer available for transcription, and the results are aggregated. The line remains legible, but greyed out.
- Once three volunteers indicate an entire document has been transcribed, it is retired from the system.

There is a large and vibrant community of researchers eager for the results of this effort, working on Lincoln, presidential and Civil War studies, submitting new words and variants to the *Oxford English Dictionary*, etc.

Full Transcription of Single-Voice Audio: The IMLS-supported Oral History in the Digital Age project[9] and subsequent efforts have signaled the need for free crowdsourcing platforms optimized for full transcription of audio collections for which automated methods fail. While GLAM audio projects have proliferated in recent years, full transcription of these collections is slow or nonexistent because of the small volunteer bases and/or the lack of funds to pay the high costs for paid transcription services. We propose to apply the lessons learned from our text transcription efforts to explore best practices in crowdsourced full audio transcription. We will constrain the two bespoke full audio transcription projects to collections featuring a single speaker, before adding the additional layers of complication required to handle speaker diarization (i.e., tagging the transcription for two or more speakers). We will also defer to a future effort tagging prosodic features of speech, e.g., stress, pitch, loudness, etc. or emotion/personality. An important feature for audio that differs from text that we will work closely with our project partners to understand is the optimal chunking of the audio clip for transcription. In text, we have found that allowing volunteers to transcribe as little as a single line at a time optimizes retention and engagement. In Chimp&See[10], we have found that 15 second video segments optimize volunteer retention and engagement while also allowing for chimp behavior identification. The optimal

[8] The challenges of crowdsourcing with TEI markup have been well described by the 'Transcribe Bentham' research team: Causer, Tim, Justin Tonra, and Valerie Wallace, 'Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham', *Literary and Linguistic Computing*, 27 (2012), 119–37 <http://llc.oxfordjournals.org/content/27/2/119.full.pdf+html> [accessed 13 December 2014]; and Causer, Tim, and Melissa Terras, '"Many Hands Make Light Work: Many Hands Together Make Merry Work": *Transcribe Bentham* and Crowdsourcing Manuscript Transcription', in *Crowdsourcing Our Cultural Heritage*, ed. by Mia Ridge (Ashgate, 2014), pp. 57–88.

[9] http://ohda.matrix.msu.edu/

[10] http://www.chimpandsee.org

minimum length of an audio clip to transcribe at a time is an open question for crowdsourced audio transcription which we will explore here.

Because full audio transcription is a new direction for Zooniverse[11], we felt it was best within the proposal writing period to identify potential partners for the two bespoke projects (see letter of support from the University of Minnesota), but not commit to any at this time. Through our open call for audio proposals at the end of Year 1, we will work closely with our advisory board to recruit, review, and make our selection.

Image Annotation/Marking: The Zooniverse has significant experience in supporting projects across the disciplines which use image annotation and marking (e.g., the Field Museum's 'Microplants' image marking project, the London Natural History Museum's 'Orchid Observers' project, amongst others). As a result, we have already built into the Project Builder image annotation and marking capabilities. We will provide training in using these tools through the proposed Project Builder workshops. We will also work directly with The Adler Planetarium Webster Institute for the History of Astronomy, home to one of the premier astronomical instrument image collections in the world, with thousands of documents from the middle ages through the 20th century. Dr. Pedro Raposo, curator for the Webster Institute, is the lead for this project. He is an ideal partner for this IMLS effort; he is a close colleague at the Adler Planetarium with PI Dr. Trouille, currently collaborating with Zooniverse on a bespoke constellation image marking project and is thus deeply familiar with Zooniverse and the opportunities the platform enables, and is committed to co-mentoring the GLAM Postdoctoral Fellow.

Through this Adler scientific instrument Project Builder built project, volunteers will first tag each image with the instrument type depicted[12], choosing among a few dozen instrument types (e.g., astrolabe, octant, orrery, sundial, sextant, telescope, etc.). A tutorial, common features to narrow down the selection, and a detailed Field Guide will provide guidance. The volunteer will then tag the image with the instrument's use. There is an active community of researchers in the history and evolution of scientific instruments, and a desire for this kind of project toolkit has been expressed by a number of other museums, including Oxford's History of Science and Ashmolean museums. Because of the variety both in type and age of instruments depicted in the Adler collection (and potential partner collections), providing an annotated and searchable database will be of great interest to this community.

Data Pipeline: The foundation for the new Zooniverse platform that includes the Project Builder is an application written in Ruby on Rails which supports a powerful Application Programming Interface (API). The API generically serves subjects – images, video or audio – for classification by volunteers via a workflow defined by the project, and receives and records these classifications into a database. The front-end Javascript software presents user interfaces to volunteers and supports the Project Builder. In order to scale dynamically to handle high traffic, the system is hosted via a Docker container on the Amazon Web Services (AWS) platform.

Data Ingestion into Zooniverse: In the current Project Builder, research teams can upload batches of 500-1000 subjects (images, videos, or audio clips) at a time by simply dragging/dropping the files. For larger collections and for bespoke projects, typically the research team provides a hard drive and the Zooniverse team uploads the subjects to the API. Through the projects proposed here, we would create a system to better support direct ingestion of large subject sets through a user-friendly web interface, adding functionality to the foundation we already have in place within the Project Builder. The current interface has three major failings: users are unable to 1) do batch uploads, i.e., set up a series of uploads in order to automate the upload of their full data set, 2) view the status of their upload, i.e., what has been uploaded and what is in the process of uploading, and 3) troubleshoot if a given subject or set of subjects need to be deleted from an upload. Through this effort we will address these issues and facilitate data ingestion into Zooniverse. This effort will lay the foundation for future work with CMSs to develop plug-ins to ingest subjects directly into the Zooniverse API.

[11] We note that through Bat Detective and WhaleFM, the Zooniverse has gained experience processing and ingesting audio files into our platform, but not in audio transcription.

[12] Done using the 'filter' function first used in Snapshot Serengeti for 72 animals and now used by over a dozen Zooniverse projects.

Useful Output for Curation: One of the major successes of the Smithsonian Transcription Center[13] approach is that the output is easily ingested in their CMS. Current Zooniverse transcription projects are not set up with this functionality. Currently, through our Project Builder for image annotation/marking projects, research teams can download the raw classification results (i.e., all classifications by all volunteers) as well as automatically-generated aggregated results that include confidence measures on consensus. Through this proposed effort, we will design data outputs for full text transcription and full audio transcription that are suitable for ingestion into different GLAM CMSs. A key aspect of this effort is to continue exploring best practices and approaches for transcription aggregation and confidence metrics, building on our efforts with AnnoTate, Shakespeare's World, etc. Our advisory board, bespoke project partners, and feedback sessions participants will guide us in this effort.

Addressing Research Q2 (Transcription with and without knowledge of others' efforts): Through the two bespoke text transcription projects, we will investigate the impact on transcription quality and volunteer experience when volunteers transcribe in isolation versus with knowledge of how others have transcribed the same document. As discussed above, Zooniverse uses a radically different approach to crowdsourcing in which multiple people perform the same task independently and the results are aggregated. We will not have volunteers edit or verify previous transcription efforts. Rather, we will examine the impact on transcription quality of volunteers being able to view other volunteers' transcriptions as well as the aggregated transcription result and how or if this new dimension changes volunteer experience, specifically engagement and retention.

In terms of measuring impact on transcription quality, we will compare the rate of accuracy for individuals who transcribe in isolation versus individuals who see previous transcriptions. We will also compare the rate of accuracy in aggregated results for lines transcribed only by those working in isolation versus for lines in which all but the first transcriber sees previous transcriptions. In terms of impact on volunteer experience, we will analyze the user behavior statistics we gather, e.g., number of transcriptions completed in a given session, length of session, number of sessions overall, sentiment analysis of discussion forum comments, etc.

There are numerous open questions in this experiment: Does knowledge of other individuals' or collective transcriptions lead individuals down the wrong path? Is transcription more or less accurate if people work in isolation or with an awareness of other people's work? Does making transcriptions visible increase retention as a result of highlighting that an individual's effort is part of a broader community effort or have the opposite effect? What environment best promotes skills acquisition, i.e. improved paleography?

Addressing Research Q3 (Feedback/Training): We will provide numerous opportunities for input and feedback from and training for the GLAM community, specifically by working closely with our advisory board and four GLAM project partners throughout and through the feedback sessions in Year 2 and the training workshops in Year 3. See Section 2 for the metrics we will use to analyze the impact on the GLAM community.

Dissemination of Results: Through this proposed effort we will produce two peer-reviewed publications. One article will focus on the methodology for creating, processing, and evaluating the data produced through this approach to crowdsourcing GLAM collections and will be submitted to the *Journal of Open Humanities Data*, on which Dr. Van Hyning sits on the Editorial Board. The second will focus on Research Q2 and will be submitted to *D-Lib Magazine*[14] or *Digital Humanities Quarterly*. In Year 3 we will present the results of our efforts at the conferences listed above, in addition to running the training workshops at those conferences. Meanwhile, all Zooniverse code is openly available on Github (https://github.com/zooniverse/).

4. PROJECT RESOURCES: PERSONNEL, TIME, AND BUDGET

Key Project Staff Effort and Duties: The Zooniverse is supported by software developers, educators, researchers, and program managers at the Adler Planetarium and the University of Oxford. This transatlantic

[13] https://transcription.si.edu/
[14] http://www.dlib.org

partnership enables Zooniverse to respond to the needs of research and GLAM communities around the world. We note in particular that both the Adler and Oxford have housed Zooniverse developers since 2007, with the team split evenly between the two institutions. This takes advantage of funding opportunities, the talent pool in both locations, proximity to project partners in the U.S., UK, and Europe.

IMLS funds will support four Adler staff: a GLAM postdoc (100% FTE for 3 years), one front-end developer (100% FTE for Years 1-2 and 1 month in Year 3), one back-end developer (73.5% FTE for Years 1-2 and 1 month in Year 3), and a web designer (100% FTE for Years 1-2) and one Oxford staff: one front-end developer (100% FTE for Years 1-2). Though no cost-share is required, Adler institutional funds will support: PI Dr. Trouille (5% FTE for 3 years), Adler lead curator Dr. Pedro Raposo (8% FTE for 3 years), and Adler administrative support. Prof. Chris Lintott and Dr. Van Hyning, Oxford Zooniverse PI and Zooniverse Humanities PI, respectively, will continue to coordinate Zooniverse effort at Oxford, and will consult on project planning and progress.

Drs. Trouille, Lintott, and Van Hyning will work closely with each other and the teams described below to provide guidance and leadership on all aspects of the project. The tables below describe the organization of our personnel into project teams and the team's responsibilities and goals for each of the three years:

Research team	GLAM Postdoc, Raposo, Trouille, Lintott, Van Hyning
Dissemination team	GLAM Postdoc, Van Hyning, Raposo
Development team	Adler and Oxford front and back-end developers and designer

Year 1:

Team	**Responsibility**
Dissemination	Open call for text bespoke projects (start of Year 1) and for audio bespoke projects (end of Year 1): advertising/recruiting and project selection.
Research	Consult with the text bespoke project partners on the interface needs and data pipeline. Refine experimental design for transcription interface. Refine the impact assessment plan, instruments, and begin data collection.
Development	Build and support the image annotation Project Builder project, the 2 text bespoke projects, and the data pipeline effort.

Year 2:

Dissemination	Host feedback workshops on bespoke projects for Project Builder planning.
Research	Consult with the audio bespoke project partners on the interface needs and data pipeline. Analyze results of transcription interface experiments. Collect and analyze performance and participation metrics.
Development	Build the 2 audio bespoke projects. Implement a small number of design templates research teams will be able to choose from in the Project Builder, depending on the project type. Build successful tools/approaches into the Project Builder. Provide ongoing maintenance for the bespoke projects.

Year 3:

Dissemination	Develop, recruit for, and run training workshops in using the Project Builder.
Research	Continue to analyze transcription interface experiments and performance and participation metrics. Write and submit articles.
Development	Provide ongoing maintenance for the Project Builder and bespoke projects.

Advisory Board: Our advisory board will provide both formative and summative guidance and feedback through quarterly virtual meetings throughout the 3-year effort. We will share an agenda and relevant review documents prior to each meeting. The board's recommendations will be used in selecting among the proposed projects in the open call, advising on our data pipeline strategy for getting crowdsourced GLAM data into catalogues, identifying successful tools and processes that emerge from the custom projects that should be incorporated into the Project Builder, and reviewing results from the research analysis. The following individuals and organizations are members of our Advisory Board:

- Dr. Meghan Ferriter, Smithsonian Transcription Center, D.C., Project Coordinator
- Rachel Frick, Digital Public Library of America, Director, Business Development
- Trish Rose-Sandler, Biodiversity Heritage Library, Data Analyst and 'Science Gossip'[15] collaborator
- Ben Vershbow, New York Public Library, Digital Labs, Director; Scribe-Zooniverse[16] collaborator

Project Finances: Dr. Trouille, with support from the Adler finance and administrative team, will manage the project finances for the Adler. The project finances under the Oxford subcontract will be managed by the Department of English's finance team, and by Dr. Van Hyning, with assistance from Prof. Lintott.

Facilities, Equipment, and Supplies: The software developers, designer, postdoctoral fellow, and guidance team have access through the Adler and Oxford to the computing equipment, office space, Internet access, and programming and design software necessary for this project. The Adler and Oxford also provide the required server space to store and manage the data and to perform the computational tasks.

5. COMMUNICATIONS PLAN:

We will communicate our efforts to GLAM institutions through advertising and recruiting for the open calls in Year 1, feedback sessions in Year 2, and the Project Builder training workshops and monthly calls in Year 3. The performance and participation metrics we will collect, analyze, and share with the GLAM community are described in detail in Section 2 above. We will also disseminate the results of our research through research publications and conferences as described in the 'Dissemination of Results' section above.

More broadly, Zooniverse projects disseminate results through social media (blogs, Twitter, Facebook), each project's discussion forum, and direct emails to volunteers. We track engagement through Google Analytics across all of our projects, including tracking the response rate to these social media posts and emails.

[15] http://www.sciencegossip.org

[16] https://github.com/zooniverse/scribeAPI

Year 1: Experimenting with Text and Image Bespoke Projects + Data Pipeline

Open call for the 3 additional bespoke projects*
- Advertising/recruiting
- Project selection

Consult with the 2 text bespoke projects partners
- Refine deliverables/scope for the pre-identified project
- Refine deliverables/scope for the open call project

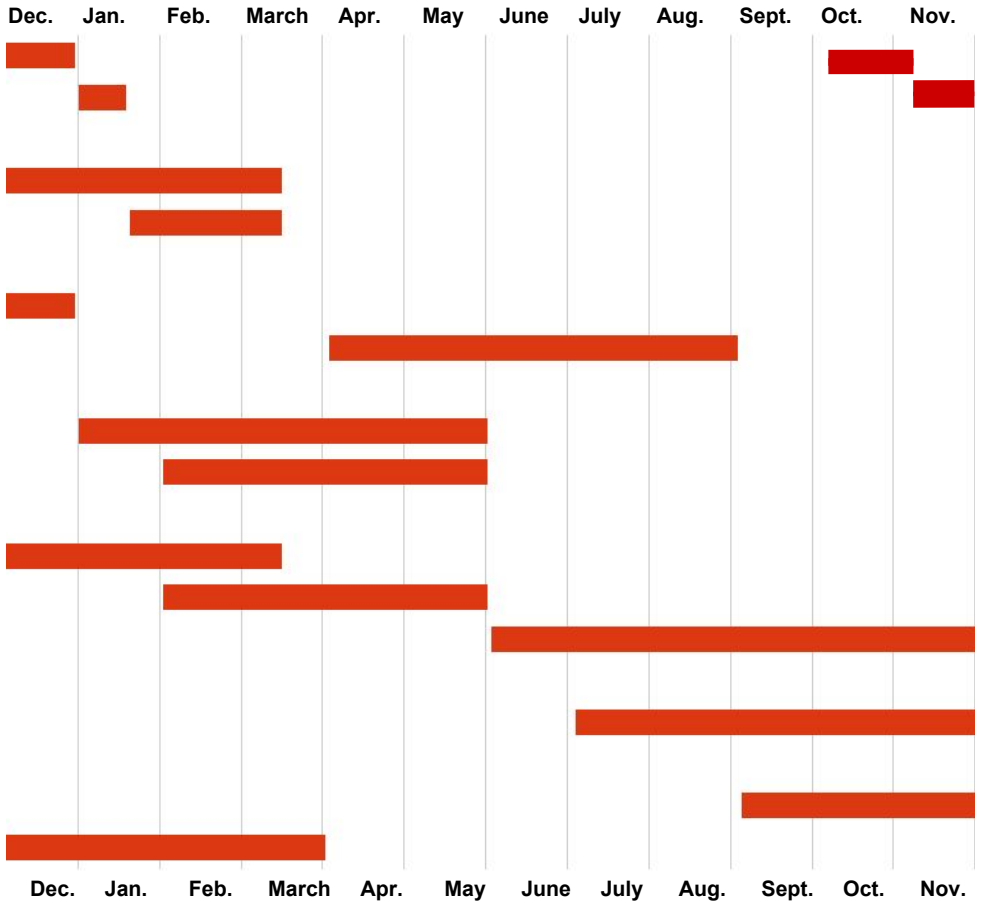
Project Builder image annotation/marking
- Minor updates to existing Project Builder annotation/marking tools
- Create Adler Scientific Instrument Project Builder built project

Bespoke Text Transcription Interface
- Revision/expansion of current transcription tools
- Tools to support transcribing with knowledge of others

Data Pipeline
- Identify data formats needed
- Support automated ingestion of data into Zooniverse
- Support useful output for curation

Deploying & Running the Custom Projects
- Launch and maintenance

Research
- Preliminary analysis: transcription w/ & w/o knowledge of others
- Create instruments for performance and participation metrics



*Note: text open call in Dec. 2016, audio open call in Nov. 2017.
We already have 1 text bespoke project (Lincoln Papers) identified.

Year 2: Experimenting with Audio Bespoke Projects + Research + Building Tools into the Project Builder

Consult with the 2 audio bespoke projects partners
- Refine deliverables/scope

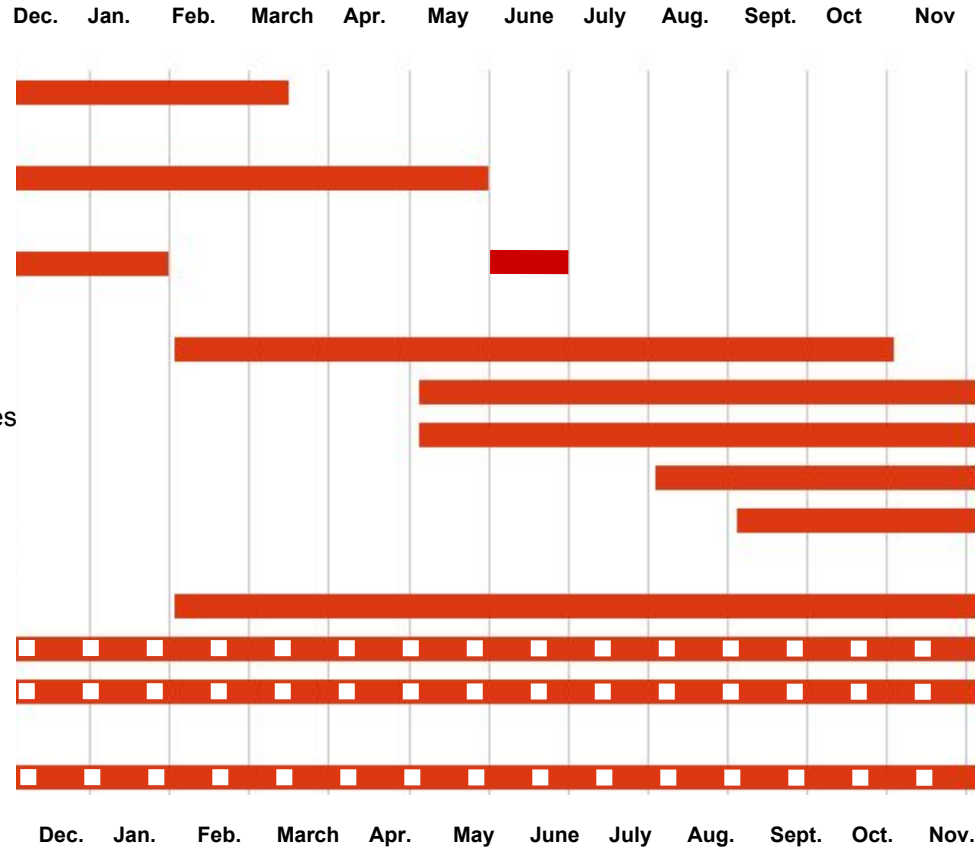
Bespoke Audio Transcription Interface
- Expand/adapt tools from Text Transcription Interface

Feedback on Custom Projects for Project Builder planning
- Host feedback workshops/sessions at conferences

Project Builder
- Build successful tools and processes into Project Builder
- Build successful data pipeline processes into Project Builder
- Build user-friendly Project Editor for new features & design templates
- Pilot test new features w/ select partners
- Revise new features

Research
- Analysis: transcription w/ and w/o knowledge of others
- Collect & analyze performance metrics
- Collect & analyze participation metrics/impact of workshops

Project Support/Maintenance
- For both Bespoke & Project Builder built projects



Indicates an intermittent activity

Months (December 2016 through November 2017)

Year 3: Research + Training/Support in using the Project Builder

Training in using the new GLAM tools in the Project Builder
- Develop training workshops

- Recruit for and run training workshops

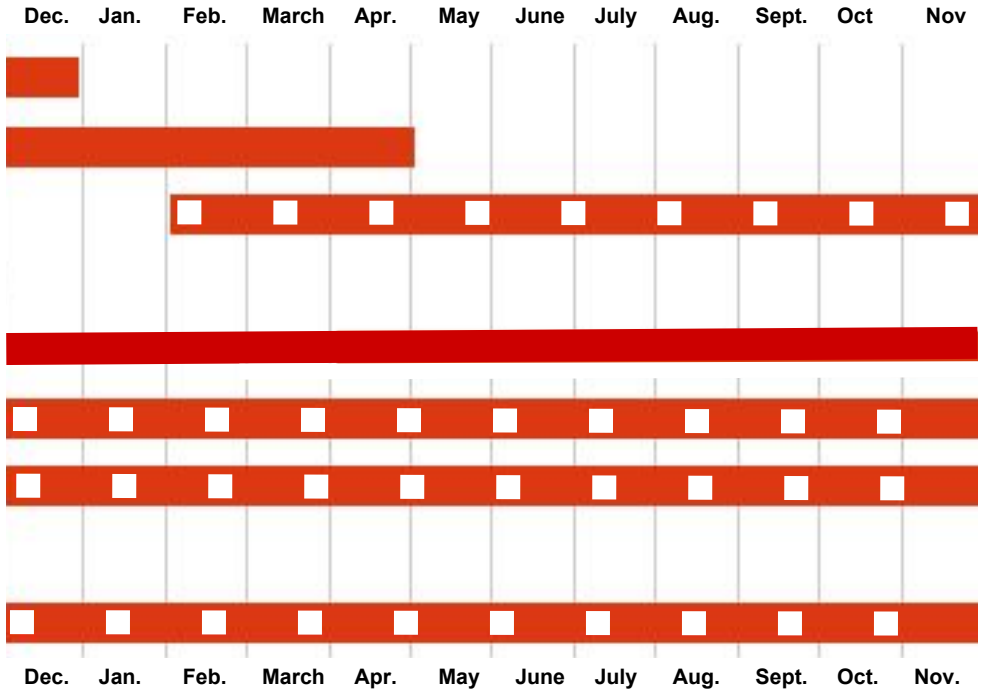
- Host monthly calls for Project Builder questions/support

Research
- Continue analyzing results; write and submit articles for review

- Collect & analyze performance metrics

- Collect & analyze participation metrics/impact of workshops

Project Support/Maintenance
- For both Bespoke & Project Builder built projects



Indicates an intermittent activity

Months (December 2016 through November 2017)

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☐	Digital content	Part II
✔	Software (systems, tools, apps, etc.)	Part III
✔	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

Zooniverse code is released under an Apache 2.0 license and is made available as open-source code on Github. It can be reused, forked and adapted without limits. Typically, all raw and aggregated Zooniverse data, generated by volunteer crowdsourcers, is released for reuse by anyone wishing to access it. Different Zooniverse bespoke project partners take different approaches to how they store and present this data during and after a Zooniverse project has run its course. Because Zooniverse is not the final repository for any data or datasets, we encourage best (meaning open and permissive) practices on the part of our research/GLAM partners and offer advice on how best to share data.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

None.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

Any data (images, audio, video) placed on the Project Builder and made public can be reused by other GLAM or research groups. It is possible that a bespoke GLAM partner could have a collection containing sensitive material or which might require obtaining permissions or rights (as was the case for the Zooniverse bespoke project AnnoTate with Tate Museums and Archives). However, as much as possible, we prefer to work with bespoke partners who follow best (open and permissive) data practices.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

The new Zooniverse platform, Panoptes, is an application written in the common Ruby on Rails framework, which supports a powerful Application Programing Interface (API). The API generically serves subjects—images, video or sound—for classification by volunteers via a workflow defined by the project. The front-end software, written using modern Javascript libraries such as React, presents user interfaces to volunteers and provides a project building tool, the Project Builder, for researchers/GLAM professionals. Six bespoke new front-end websites will be built in React and interact with the Panoptes API. The six bespoke projects will provide a foundation for new tools that will be built into the Project Builder. The primary audience(s) the software built specifically through this proposed effort will serve are GLAM institutions and their crowdsourcing volunteers.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

Other platforms for citizen science and crowdsourcing exist, most based on a radically different model from the volunteer-driven collaborative environment of the Zooniverse. Amazon's Mechanical Turk, for example, pays volunteers to carry out microtasks, a model which is not sustainable for most GLAM/research teams. Crowdcrafting.org, powered by PyBOssa, provide online tools for project building, but their community is much smaller than ours and their support for research teams is less advanced. Crowdcrafting.org also lacks a consistent approach to aggregating or validating data.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

Zooniverse sites and tools are built using JavaScript, often React, and a Ruby on Rails application. These technologies make best use of the capabilities of modern browsers and provide a smooth user-experience. The data – whether in the form of subjects (images, audio or video to be classified), classifications performed by volunteers, or interactions in discussion tools – are stored in a PostgreSQL database. This format allows data to be stored alongside rich and varied metadata.

B.2 Describe how the intended software will extend or interoperate with other existing software.

New front-end web apps built in React will communicate with the Zooniverse Panoptes API, thus reaping the advantages of that ecosystem, while also implementing new tools and experimental approaches to the problems of text transcription, audio transcription, and guided investigation. Successful and robust tools and approaches will be incorporated into the Project Builder.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

The Panoptes system is hosted via a Docker container on the Amazon Web Services (AWS) platform. AWS is a cloud environment, which supports a variety of services including database storage, virtual hosting, and a queuing system for commands. As a USA-UK based organization, we benefit from the UK JISC agreement for bulk hosting on AWS. As well as increased scalability and flexibility, this saves on the cost of local hosting, while retaining future flexibility.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

All Zooniverse code is available as open-source code on Github. We will continue to update and document this code base for wider use. As new features and tools are added to the Project Builder, existing Zooniverse 'best practices' and 'how-to-build-a-project' documents will be updated and expanded.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

~35 projects were built on the old Zooniverse API called Ouroborus. These include http://penguinwatch.org/; http://www.sciencegossip.org/ and http://www.operationwardiary.org/, to name only a few. More recently, custom front end projects built in Angular.js were created on the Panoptes API, including AnnoTate (https://anno.tate.org.uk/#/) and Shakespeare's World (http://www.shakespearesworld.org/#/). A recent NEH-funded collaboration with NYPL led to Scribe https://github.com/zooniverse/scribe and a number of new projects, e.g. http://www.measuringtheanzacs.org/. See https://www.zooniverse.org/ for the Zooniverse homepage and https://www.zooniverse.org/lab/ for the Project Builder.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

Zooniverse code is released under an Apache 2.0 license and is made publicly available on Github. It can be reused, forked and adapted without limits.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

All Zooniverse code is available on Github: https://github.com/zooniverse. Some repositories are closed while under active development, but are all ultimately made freely available. Panoptes code is available here: https://github.com/zooniverse/panoptes (API) and Project Builder code is available here: https://github.com/zooniverse/panoptes-front-end. All bespoke projects which communicate with the API have their own front-end repository, i.e. https://github.com/zooniverse/shakespeares_world.

C.3 Identify where you will be publicly depositing source code for the software developed:

Github

Name of publicly accessible source code repository: Github/Zooniverse
URL: https://github.com/zooniverse

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

Crowdsourced metadata is produced by multiple volunteers assessing the same objects independently. Their responses are then aggregated using clustering and other aggregation techniques drawing on, for example, hidden Markov models. Aggregation code and data cleaning methodologies are posted to https://github.com/zooniverse/aggregation. Volunteer classifications and discussion on Talk are saved as JSON in a PostgreSQL database, which maximizes flexibility by avoiding the need to define a rigid structure for classifications, while still providing easy search. A csv output of metadata can be requested at any time under the 'data export' tab in the Project Builder for bespoke and out of the box projects, as can a full export of 'Talk' discussions. Zooniverse will work closely with the advisory board and bespoke project partners to ensure these projects gather necessary metadata in a form that is most useful to the GLAM community.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

No.

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

The Zooniverse holds a small volume of personal data necessary to carry out site operations, primarily email address and (if given) real name, along with a record of IP addresses, browser, hardware and operating system information. This data is held securely on Amazon Web Services' data store (a signatory to the US/EU Safe Harbor agreement) and is not distributed to researchers. All Zooniverse user data is made anonymous and volunteers are only referred to by user ID in publications unless explicitly asked for permission to be referred to by name.

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

N/A

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

The Zooniverse software is described in Section III above.

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

The code (all open source) for the bespoke projects will be stored at github.com/zooniverse. Documentation is included as supplements to the codebases. Once we have implemented new GLAM tools and approaches into our Project Builder, we will update our Project Builder documentation for how to use these specific tools, as well as how to aggregate and ingest the crowdsourced metadata from these tools. The documentation is shared via websites linked to the Project Builder.

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Metadata generated by each project will ultimately reside within the GLAM CMSs and, as appropriate, in additional repositories such as the Digital Public Library of America (DPLA) and the Database on Demand Service hosted by CERN. Zooniverse policy is to preserve project crowdsourcing websites and their data even after their main goals are achieved, as this enables further educational use as well as providing context for datasets collected using them (see, for example, http://old.oldweather.org). These sites are currently hosted on AWS and there is funding to support zooniverse.org until 2020. If this level of support becomes unavailable, it would be possible to transfer to a free hosting service (e.g. Heroku). Although not suitable for production sites, this would be satisfactory for archiving. The release of open source code via GitHub also ensures continued and sustained access.

8. Identify where you will be publicly depositing dataset(s): AWS, DPLA, Database on Demand

Name of repository: AWS, DPLA, Database on Demand
URL: aws.amazon.com, http://dp.la, http://information-technology.web.cern.ch/services/database-on-demand

In particular, Zooniverse is keen to find additional ways of sharing data via the Digital Public Library of America. Dr. Van Hyning (Oxford co-I) is on the editorial board of a new journal devoted to data dissemination, The Journal of Open Humanities Data, which enables authors to describe their datasets, data cleaning and storage strategies, and make suggestions for further research using that data.

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

At the start of the grant period, we will review and update our data management plan in close consultation with our advisory board. In collaboration with our six bespoke project partners, we will create a detailed data management plan for the metadata created by each project (with a priority set on ultimately making the metadata easily accessible to the public). Through the feedback sessions in Year 2 and in further consultation with our advisory board, we will review and update our data management plan for GLAM Project Builder built projects.

Original Preliminary Proposal

IMLS Research Grant: 'Transforming Libraries and Archives through Crowdsourcing'

As libraries, museums, and other cultural repositories digitize their collections and place them online, the challenges of transforming these materials into useful and searchable sources of information are becoming increasingly apparent. While OCR and handwriting recognition technology have opened up some print and manuscript corpora, and image recognition software is improving daily, there are still many tasks that require human intervention. For these, volunteer crowdsourcing is a viable and vibrant solution.

Crowdsourcing projects have proliferated in the last decade, but there are, as yet, no agreed upon standards or platforms—and particularly open source and free platforms—on which galleries, libraries, archives, and museums (GLAM), academics, and amateurs can rely. Zooniverse, the world-leading academic crowdsourcing organization, has the technical expertise to intervene at this critical stage and offer new, free, and reliable tools. Since 2007, Zooniverse has attracted over 1.34 million volunteers; collaborated with more than 500 GLAM and academic professionals around the world, and created over thirty-five online projects, including five in GLAM and the humanities (e.g., 'Operation War Diary' [operationwardiary.org] and 'AnnoTate' [Anno.Tate.org.uk]). Zooniverse projects have resulted in over 100 scholarly publications and numerous discoveries. The Zooniverse is unusual among crowdsourcing organizations in the degree to which it fosters user-centric design, specialist-amateur dialogue, as well as amateur research and publication. This firmly fits within the IMLS' remit to place users at the center of the national digital platform because 'they are ultimately our strongest advocates'.

Research focus:
Zooniverse is poised to serve as a transformative element in the future of GLAM research efforts and the national digital platform envisioned by the IMLS. To best serve the GLAM community, we need to understand what crowdsourcing tools and data outputs would be most useful for opening up GLAM corpora. In order to build that understanding, through this proposed effort we will address the following research questions:

- Q1: How can crowdsourcing be deployed in the arenas of metadata extraction and text transcription for the purposes of unlocking big data currently trapped in GLAM sources that cannot be machine read?
- Q2: How does synchronous versus asynchronous collaboration by users in GLAM crowdsourcing projects impact the quality and depth of analysis and participation?
- Q3: How can we extend our crowdsourcing expertise to more librarians and archivists and learn from them, in turn, how to adjust the Zooniverse platform to best meet their research needs?

Addressing Q1/Q2/Q3: Open call for new GLAM Zooniverse projects:
Zooniverse is regularly approached by libraries and archives from the U.S. and abroad (~6 per week) who have viable and interesting projects, but we do not currently have the time or sufficient in-house expertise to identify those that are the best for development. If awarded an IMLS grant, we will build 6-8 custom-made GLAM projects, identified through an open call. New projects could focus on text, object, image, video, or audio sources. We will appoint a project lead with a library/archives background to lead this effort, working closely with our GLAM partners and Zooniverse technical staff. While we have expertise in building tools for science, we need to work closely with GLAM partners–hence the custom-made projects–in order to truly understand what tools and data outputs are most useful. These new projects will allow us to experiment with different approaches to metadata extraction and text transcription; for example, comparing the transcription quality between users transcribing a single line versus a whole page in a given session. We will also experiment with

synchronous versus asynchronous collaboration approaches and interfaces (e.g., comparing the metadata extraction results between users working independently, working together synchronously, and working together but asynchronously). These research results will provide an important contribution to the field, to improving the discoverability and functionality of GLAM digital content, and in guiding design choices for future GLAM and humanities crowdsourcing projects.

Addressing Q3: Supporting GLAM organizations in building their own crowdsourcing projects:
In 2015, Zooniverse created a unified platform, Panoptes, which allows researchers from around the world to build their own crowdsourcing projects for free, with no computer programmer support. Prior to Panoptes, each new Zooniverse project required funding to support significant developer effort. Through the custom-made projects discussed above, we will identify and refine tools required by GLAM crowdsourcing projects. As part of this effort, we will incorporate refined tools into our Panoptes platform, available for free for all to use. The refined tools for GLAM crowdsourcing projects in Panoptes might include:

- text tagging and transcription crowdsourcing tools for GLAM corpora
- metadata extraction from photographic, video, audio, or other resources in GLAM collections, building on our efforts in the sciences; e.g., processing whale sounds and chimpanzee videos.
- the ability for users to work collaboratively, both synchronously and asynchronously.

At the end of years 1 and 2 of this effort we will host focus group workshops. The primary goal of the year 1 workshop will be to obtain feedback from GLAM practitioners on the approaches and tools used in the custom-made projects and guidance on what to incorporate into Panoptes. The primary goal of the year 2 workshop will be to recruit and train new GLAM practitioners in using these crowdsourcing tools.

Advisory Board:
Our advisory board will provide guidance and feedback throughout the three-year effort, including helping our team to choose among proposed projects and advising on a long term data strategy for getting crowdsourced data into catalogues. The following individuals and organizations are on board:

- Ben Vershbow, New York Public Library, Digital Labs, Director; collaborator on the Scribe-Zooniverse codebase: https://github.com/zooniverse/scribeAPI
- Dr. Meghan Ferriter, Smithsonian Transcription Centre, D.C., Project Coordinator.
- Trish Rose-Sandler, Biodiversity Heritage Library (BHL), Data Analyst and collaborator on the BHL-Zooniverse project Science Gossip (sciencegossip.org).

Budget:
The total budget for this proposal is $1.36M over three years ($1.27M grant-funded and $94K in cost share). Grant funds support the salary and benefits for four staff at the Adler Planetarium: librarian/archivist postdoc, one front-end and one back-end developer, and a designer. Grant funds also support the salary, benefits, and indirect costs of two staff at University of Oxford through a subaward: front-end developer and GLAM Project Manager. Grant funds also support travel and web hosting costs. Adler's indirect costs are included per our NSF negotiated rate. Although not required in IMLS Research Grants, cost share is provided through institutionally supported partial salary and benefits and associated indirect costs of three staff at the Adler: Director of Citizen Science, Technical Lead, and Archivist & Librarian. Additional cost share from Oxford is provided through partial salary and benefits and indirect costs of the Zooniverse PI.