PROCEEDINGS OF SCIENCE

From Clicks to Publications: How the Public is Changing the Way We Do Research

Laura Trouille[*†]
The Adler Planetarium and Northwestern University
E-mail: trouille@zooniverse.org

Chris Lintott
University of Oxford
E-mail: chris@zooniverse.org

Lucy Fortson
University of Minnesota-Minneapolis
E-mail: fortson@physics.umn.edu

Processing our increasingly large datasets (e.g., image, audio, video, multidimensional data, etc.) poses a bottleneck for producing real scientific outcomes. Citizen science - engaging the public in research - provides a solution, particularly when coupled with machine learning algorithms and sophisticated task allocation and retirement rules. Zooniverse is the most widely used and successful citizen science platform, with almost 1.5 million volunteers worldwide and having supported over 50 projects across the disciplines (including ecology, archaeology, climate science, oncology, physics, astronomy, and the humanities). To date, Zooniverse projects have produced over 100 peer-reviewed publications. Faced with a rapidly growing demand for citizen science projects, Zooniverse has launched a new 'Project Builder' interface which allows anyone to build and maintain their own citizen science project using a set of browser-based tools.

BASH 2015
18 - 20 October, 2015
The University of Texas at Austin, USA

[*]Speaker.
[†]A sincere thank you to our volunteers who make the Zooniverse possible.

PoS(BASH2015)009

1. Overview

Since the launch of Galaxy Zoo[1] in 2007, the Zooniverse[2] has grown into the world's largest and most successful platform for citizen science online. With a workforce of nearly 1.5 million registered users (more than half of whom have been active in the last year) providing $\sim 100,000$ classifications every day, an email to the Zooniverse volunteers is a powerful recruiting tool for any new project. At a time when citizen science is gaining in prominence across the globe, the Zooniverse is a core part of the research infrastructure landscape.

Zooniverse has now supported more than fifty projects over a wide range of research domains, producing data that has been used in more than a hundred peer reviewed papers. For example, in ecology, the Snapshot Serengeti project has used camera trap image classifications to estimate the density of lions in the Serengeti National Park, Tanzania [2] and in meteorology the Cyclone Center project allows for better characterization of global climate change through studying hurricane strength [3]. A multitude of humanities projects utilize transcriptions made by the public, including the Ancient Lives project that transcribes ancient Greek on fragments of papyri [12]. There are also many examples of Zooniverse projects in astronomy, from the SpaceWarps[3] gravitational lens discovery project [6] to identifying protoplanetary disks in Disk Detective[4]. A full review is given in [7], which highlights the twin key advantages of this approach: combining the ability to scale analysis to the size of modern datasets with the ability to make serendipitous discoveries. This ability of the crowd to react to the unusual and interesting is a key advantage, and has resulted in a variety of important discoveries across many projects, from AGN-ionised gas clouds like 'Hanny's Voorwerp' [5], to the only planet known to date in a four-star system [10]. Zooniverse projects support a range of data types (including image, video, audio, plots, and text) and tasks (including tagging, marking/drawing, and text transcription). The Adler Planetarium and the University of Oxford house the core development and communications team, in collaboration with the University of Minnesota-Minneapolis and the broader Citizen Science Alliance.

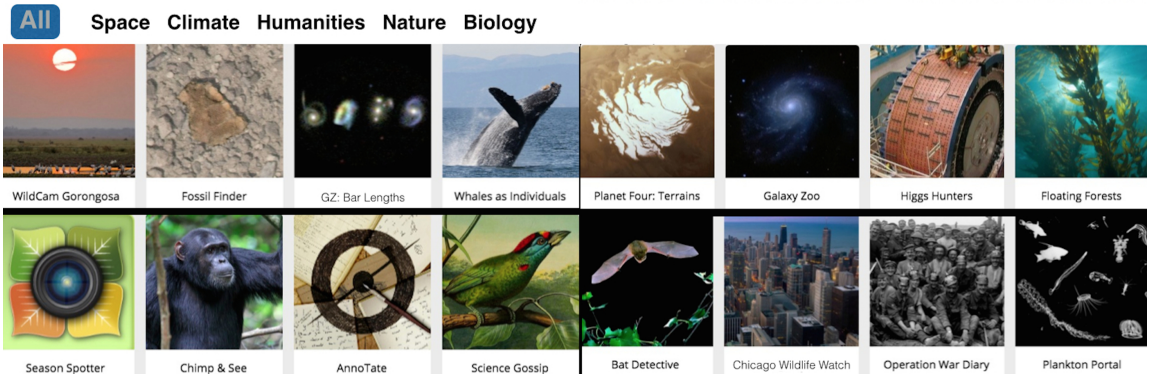

Figure 1: A selection of the over 50 Zooniverse projects the public have participated in to date.

[1] http://www.galaxyzoo.org
[2] http://www.zooniverse.org
[3] http://www.spacewarps.org
[4] http://www.diskdetective.org

2. DIY Zooniverse: The Project Builder

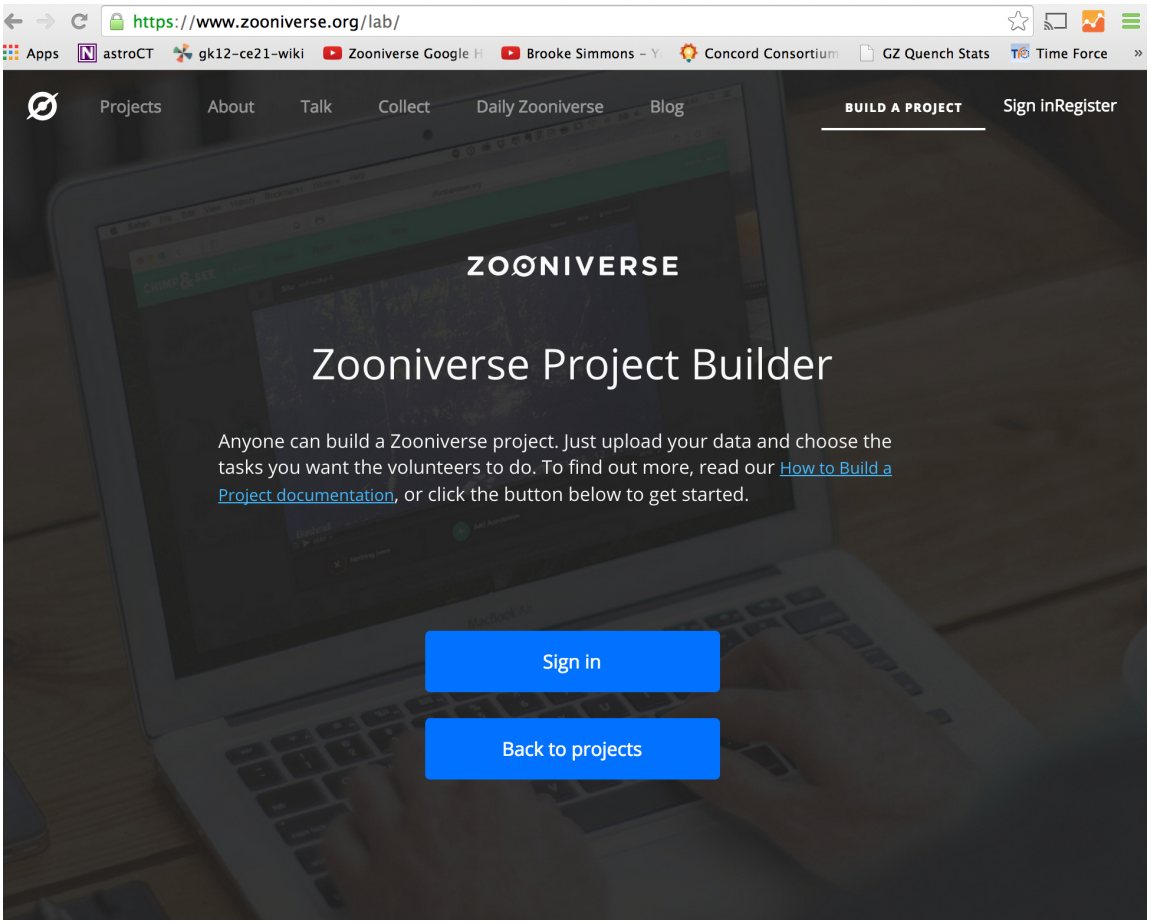

Figure 2: The Zooniverse Project Builder interface enables anyone to build and maintain their own citizen science project using a set of browser-based tools.

The willingness of millions of volunteers to sort through modern datasets allows analysis at scale, with the opportunity not only for routine categorization but also for serendipitous discoveries. However, access to this crowd has, until now, relied on researchers having the technological facility to build modern, responsive web sites with excellent user interface design and a good understanding of the twin problems of volunteer motivation and efficient task allocation. The Zooniverse solves this problem through its new Project Builder platform, which launched in June 2015. The platform enables researchers to build and maintain their own citizen science projects using a set of browser-based tools. Prior to its development a typical project required professional web development, and even rapid projects took months to develop. With the Project Builder, anyone can build and deploy a project in an hour. The Project Builder supports the most common types of interaction, including classification, comparison, marking and drawing tools, or any combination. In dramatically reducing the effort needed for project deployment, this system makes it possible to use citizen science for smaller projects. It also enables large teams working with proprietary data to use the platform in-house. Furthermore, while raw classification tables are made available, the

use of a common platform enables the Zooniverse team to provide 'standard' aggregations of data, deriving both classifications and confidence measures from the volunteers' input. Also, project owners can invite collaborators and tag them as 'experts' (whose classifications are included in the gold standard data), 'researchers', 'moderators', and 'testers'. And every Project Builder built project has an associated 'Talk' or discussion forum in which volunteers can engage with each other and the research team.

In the first four months of the Project Builder's existence, over 700 people/teams started a project, with a half dozen already publicly promoted through Zooniverse.org. Every week new Project Builder built projects are submitted for beta review (in which Zooniverse volunteers and team members provide feedback on the content and workflow). As one example, the fossilfinder.org site, which launched in September 2015, was built entirely using the Project Builder and within the first few weeks had identified sites of interest for the team hunting hominid fossils. A major focus of the redesign was the ability to run real-time and near real-time projects with rapid data analysis, a facility utilized both by supernova-hunting astronomers and by rescue teams responding to the earthquake in Nepal.

2.1 Details on the Zooniverse Infrastructure

The software requirements for citizen science online are stringent. The interface to the data must be easily understood by a wide range of volunteers. Also, while a steady background rate of classifications is achieved by most projects, traffic can be extremely uneven when media drives traffic to a project. For example, BBC?s Stargazing Live program featured Snapshot Supernova, which then received more than a million classifications in under half an hour. In such circumstances, being able to scale rapidly and prioritizing recording classifications is essential. In order to achieve the required scaling, the system is hosted via a Docker container on the Amazon Web Services (AWS) platform.

The platform is also easily extendable, with a design that allows for iterative improvement of features in order to support the large range of projects. This has led to the adoption of a modular approach to software system design that underlies the custom-built system that supports new Zooniverse projects. At its core, the system is an application written in the common Ruby on Rails framework, which supports a powerful Application Programming Interface (API). The API, upon request, generically serves subjects (e.g., images, video, audio, etc.) for classification by users via a workflow defined by the project. The workflow, which can be a single task or an ordered set, is chosen based on the properties of the subject, so that subjects in Galaxy Zoo from different surveys, for example, can have different workflows. A task can be as simple as a single question, or in principle as complicated as a 3D model. The core functions of the API are allowing the creation and management of projects, passing subjects to the front-end software for classification, and the receiving and recording of classifications. Further modules handle secondary tasks such as data aggregation, discussion, or the provision of statistics on project progress or user behavior. Data is saved in a format (JSON in a PostgreSQL database) that maximizes flexibility by avoiding the need to define a rigid structure for classifications, while still providing easy search. Classification data is stored in a database, with registered volunteers identified only by a randomly assigned ID. Personal data on classifiers obtained by the Zooniverse is limited to email address and (optionally) a full name, but this is stored in a separate secure database. Volunteers contributing to projects

without logging in are identified by their IP address, which is not recorded for logged in users ensuring anonymity is maintained. The front-end software, written using modern Javascript libraries, presents user interfaces to volunteers and supports the Project Builder discussed in the previous section. .

3. Engaging our Volunteers

As of Oxtober 2015, the Zooniverse has almost 1.5 million registered users from 237 countries. Approximately one-third of participants are in the United States, one-third in the United Kingdom, and the remaining third are located around the world. Zooniverse volunteers put in 216,000 hours of effort between September 2014 and September 2015. This is equivalent to 108 people working full-time (40 hour weeks) for the year.

The main reasons Zooniverse volunteers listed in a survey for why they participate in Zooniverse are:

- I like to contribute to scientific progress ($\sim 90\%$)
- I find completing the tasks fun ($\sim 50\%$)
- I am fascinated by the projects I contribute to ($\sim 85\%$)
- Distraction: It keeps my mind off things ($\sim 15\%$)
- Other ($\sim 9\%$)

An illustrative example of volunteers' deep interest in contributing to scientific progress comes from the Snapshot Supernova project, featured on BBC Stargazing Live in March 2015. For this project, 40,000 volunteers participated, providing 1.9 million classifications. Within the first few hours of the project launch, a new supernova was identified. Among the volunteers was one who classified tens of thousands of images; an enormous contribution in a short amount of time. Of all the classifications this volunteer provided, none were of one of the discovered supernovae. But this did not detract from his experience. In an email he sent to Zooniverse, he wrote, 'I posted a thank you on the forums but would like to reinforce how thankful I am. From a civilian point of view, being afforded such an amazing opportunity to contribute to an important scientific project is nothing short of incredible.'

The initial way to engage with Zooniverse projects is through the classification interface, whether tagging animals in the Serengeti Desert or classifying galaxies in Galaxy Zoo. A few Zooniverse projects have developed educational interventions within the classification interface. These 'interventions' appear as pop-up slides with information about the subject and the research field and opportunities to explore further. A few projects have created formal guides for educators on how to engage students in classrooms with Zooniverse projects and concepts (for example, the PlanetHunters educators guide[5]). The Zooniverse was recently awarded a National Science Foundation grant to develop a suite of curricular materials for the undergraduate introductory astronomy course for non-majors. The team will develop, implement, and assess the impact of an

[5]http://www.planethunters.org

authentic research experience based on the citizen science model for engagement. Students will access Zooniverse astronomical data, test hypotheses, analyze data, draw conclusions, and discuss with members of their group, the class, and the wider Zooniverse community.

All volunteers are also invited to engage more deeply through 'Talk'. Each Zooniverse project has a 'Talk' interface in which volunteers discuss questions and concepts with each other and the research team members. There is also a global Zooniverse 'Talk'. Through discussions in 'Talk', volunteers and research teams are able to explore whether potentially unusual objects are truly interesting. Some of the most important Zooniverse discoveries have been made through 'Talk'; e.g., AGN-ionised gas clouds like 'Hanny's Voorwerp' [5] and the only planet known in a four-star system [10].

A few projects have taken advantage of the 'Talk' interface to specifically invite volunteers to engage more deeply in the research topic through specific tasks. For example, in Chimp&See[6] volunteers help the research team identify (and name!) individual chimps across different videos. In Disk Detective[7], a subset of prolific volunteers meet with the research team via Google Hangouts every other week, vet potential observing targets through an in-depth analysis of metadata, and have even participated in Telescope observing runs to carry out those follow-up observations. Galaxy Zoo: Quench was the first Zooniverse-based project (and one of few citizen science projects) to support the public in experiencing the entire process of science, from galaxy classification to data analysis to writing up the results for publication in a professional, peer-reviewed journal. This experiment made clear that our volunteers are extremely eager for these deeper engagement opportunities. We expect this to be an enormous area of growth and expansion for the Zooniverse into the future, especially with the new opportunities afforded through the Project Builder.

4. Expanding the Zooniverse: The Human-Machine System

In order to respond most efficiently to the increasing data deluge across the disciplines, citizen science platforms need to be more complex - incorporating intelligent task assignment, responsive retirement rules, and machine learning strategies. A few Zooniverse projects have experimented with aspects of the above; e.g., setting different retirement rules (i.e., when a subject is removed from the system) for different sequences of classification results, taking advantage of known 'expert' volunteers in the system to more quickly retire subjects shown to these volunteers, and incorporating machine learning at various stages in the workflow. Building on this foundation, the Zooniverse is in the early stages of exploring the dynamic combination of human and machine classifiers, gaining for the first time knowledge of how load can be optimally shared in a real, flexible citizen science platform. The infrastructure under development is one that combines human classifications through an intelligent task assignment engine with machine classifications in the context of a sequence of binary tasks (sometimes referred to as cascade filtering).

A growing body of theoretical work, often using data from Zooniverse projects, has demonstrated that efficiencies exist in task assignment to volunteers that could greatly reduce the burden on classifiers [4]. Other work suggests that efficiencies can be gained through the judicious coupling of machine and human classifiers [9]. For example, presenting tasks in order of increasing

[6] http://www.chimpandsee.org
[7] http://www.diskdetective.org

machine confidence reduces the time to obtain a given target accuracy by 63% [11]. Moving away from random task assignment to a more complex model enables better integration of machine classification alongside human efforts. Until now, machine learning researchers have concentrated on using the large training sets produced by citizen science projects in order to train algorithms. These studies underline the need for extremely large training sets. For some classification problems, such training sets can easily be assembled from simulations. However, the vast majority of classification problems will face difficulties in building up a large training set; a particular difficulty when searching for rare objects. A dynamically combined machine-human system provides the best strategy. As machine learning improves, the system will allocate more tasks to it, freeing human classifiers to move on to more difficult problems. A hybrid classification system is not only potentially more accurate, but the efficiencies it enables are necessary in order to deal with modern datasets.

The Zooniverse also has the capacity through established infrastructure to conduct experiments in order to better understand what keeps the volunteer base engaged in projects. There is evidence that tasks which can be completed quickly are more popular with the volunteers, and lead to more successful projects [1]. New volunteers are often also less confident about their ability [8] and those with greater upfront learning required are less popular.

Grounded in the above research as well as the wealth of experience to date accumulated by the Zooniverse team, the Zooniverse is beginning to investigate the potential of a new type of classification system that combines the time and effort provided by millions of citizen scientists with modern machine learning. By making efficient use of both capacities, Zooniverse will achieve greater accuracy and flexibility than has been possible to date.

References

[1] J. Cox, E. Oh, B. Simmons, C. Lintott, K. Masters, A. Greenhill, G. Graham, and K. Holmes. *Defining and measuring success in online citizen science: A case study of zooniverse projects.* Computing in Science Engineering, 17(4):28-41, July 2015.

[2] J. J. Cusack, A. Swanson, T. Coulson, C. Packer, C. Carbone, A. J. Dickman, M. Kosmala, C. Lintott, and J. M. Rowcliffe. *Applying a random encounter model to estimate lion density from camera traps in Serengeti national park, Tanzania.* The Journal of Wildlife Management, 1014, 2015.

[3] C. C. Hennon, K. R. Knapp, C. J. Schreck, S. E. Stevens, J. P. Kossin, P. W. Thorne, P. A. Hennon, M. C. Kruk, J. Rennie, J.-M. Gad?ea, M. Striegl, and I. Carley. *Cyclone center: Can citizen scientists improve tropical cyclone intensity records?* Bulletin of the American Meteorological Society, 2015/01/26, 2014.

[4] E. Kamar and E. Horvitz. *Planning for crowdsourcing hierarchical tasks.* In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS, 1191-1199, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems.

[5] W. C. Keel, A. M. Manning, B. W. Holwerda, C. J. Lintott, and K. Schawinski. *The Ultraviolet Attenuation Law in Backlit Spiral Galaxies.* Astronomical Journal, 147, Feb. 2014.

[6] P. Marshall, A. Verma, A. More, C. Davis, S. More, A. Kapadia, M. Parrish, C. Snyder, J. Wilcox, E. Baeten, C. Macmillan, C. Cornen, M. Baumer, E. Simpson, C. Lintott, D. Miller, E. Paget, R. Simpson, A. Smith, R. Kueng, P. Saha, T. Collett, and M. Tecza. *Space Warps: I. Crowd-sourcing the Discovery of Gravitational Lenses.* ArXiv e-prints, Apr. 2015.

[7] P. J. Marshall, C. J. Lintott, and L. N. Fletcher. *Ideas for Citizen Science in Astronomy.* Annual Review of Astronomy and Astrophysics, 53, 247-278, Aug. 2015.

[8] G. Mugar, C. Osterlund, K. D. Hassman, K. Crowston, and C. B. Jackson. *Planet hunters and seafloor explorers: Legitimate peripheral participation through practice proxies in online citizen science.* In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing, CSCW, 109-119, New York, NY, USA, 2014. ACM.

[9] O. Russakovsky, L.-J. Li, and L. Fei-Fei. *Best of both worlds: human-machine collaboration for object annotation.* In CVPR, 2015.

[10] M. E. Schwamb, J. A. Orosz, J. A. Carter, W. F. Welsh, D. A. Fischer, G. Torres, A. W. Howard, J. R. Crepp, W. C. Keel, C. J. Lintott, N. A. Kaib, D. Terrell, R. Gagliano, K. J. Jek, M. Parrish, A. M. Smith, S. Lynn, R. J. Simpson, M. J. Giguere, and K. Schawinski. *Planet Hunters: A Transiting Circumbinary Planet in a Quadruple Star System.* Astrophysical Journal, 768:127, May 2013.

[11] A. Veit, M. J. Wilber, R. Vaish, S. Belongie, J. Davis, V. Anand, A. Aviral, P. Chakrabarty, Y. Chandak, S. Chaturvedi, C. Devaraj, A. Dhall, U. Dwivedi, S. Gupte, S. N. Sridhar, K. Paga, A. Pahuja, A. Raisinghani, A. Sharma, S. Sharma, D. Sinha, N. Thakkar, K. B. Vignesh, U. Verma, K. Abhishek, A. Agrawal, A. Aishwarya, A. Bhattacharjee, S. Dhanasekar, V. K. Gullapalli, S. Gupta, C. G, K. Jain, S. Kapur, M. Kasula, S. Kumar, P. Kundaliya, U. Mathur, A. Mishra, A. Mudgal, A. Nadimpalli, M. S. Nihit, A. Periwal, A. Sagar, A. Shah, V. Sharma, Y. Sharma, F. Siddiqui, V. Singh, A. S., P. Tambwekar, R. Taskin, A. Tripathi, and A. D. Yadav. *On optimizing human-machine task assignments.* CoRR, abs/1509.07543, 2015.

[12] A. Williams, J. Wallin, H. Yu, M. Perale, H. Carroll, A.-F. Lamblin, L. Fortson, D. Obbink, C. Lintott, and J. Brusuelas. *A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments.* In Big Data (Big Data), 2014 IEEE International Conference on, 100-105, Oct 2014.