The two-and-a-bit page guide to running a Zooniverse project.

The Zooniverse (www.zooniverse.org) is a collection of online citizen science projects. It was inspired by the success of the oldest and largest of these projects, Galaxy Zoo (www.galaxyzoo.org), which has accumulated more than 150 million classifications of galaxies, and which recently relaunched with images drawn from deep Hubble Space Telescope surveys. The guiding principle behind the Zooniverse is that sharing a common framework and volunteer base reduces the overhead in running one of these projects, making it possible to launch new projects extremely easily.

Encouragingly, a survey of roughly 10,000 Galaxy Zoo volunteers revealed that their primary self-reported motivation was to contribute to research. This suggests that there is a latent desire to help with scientific research, and indeed public response to these projects can be enormous; we estimate that the total manpower employed in Galaxy Zoo 2 alone (which ran for 14 months) was the equivalent of employing a single classifier for more than 110 years. This places a huge responsibility on us, both to appropriately credit volunteers both collectively and where appropriate and possible individually, but more fundamentally to develop appropriate projects that do not waste participants' time.

Project design

The first requirement of project design is therefore to define a data set and interaction which will lead to meaningful results. This is the responsibility of the science team associated with each of the Zooniverse projects, who must address both large-scale (e.g. which subsets of available data) and detailed issues (e.g. what zoom level is appropriate for viewing images). Our experience suggests that a large proportion of a project's classifications may come from the initial spike in interest and so it is important that the interface is correct from the start.

To help, we have compiled a questionnaire which is attached to the end of this document. It is not intended to be exhaustive, but rather to suggest some of the things that will need to be considered. While a large part of our current modus operandi is to try very different things, a typical citizen science project will involve a task for which human interaction is necessary; in other words, existing machine learning tools should not be able to complete the task to the same degree of accuracy. To take Galaxy Zoo as an example, determining morphology is an appropriate task, but measuring colour is not as that can easily be automated. The results of such a task should be structured so that an average answer can be easily established; selecting from a list of alternatives is thus in most cases more useful than free-form tagging. The task should of course be accessible to the likely volunteers (or one should be prepared to train them) – if one is going to end up searching through the dataset for the answers provided by a small number of experts, one might as well have found the experts in the first place rather than creating a citizen science task.

Once constraints on the necessary interface have been established, development can get underway. The scale of the development will be dictated by the necessary interface; it could be as simple as reusing an existing interface (for example, the Galaxy Zoo interface could be repurposed for any classification of images via a decision tree) or as

complicated as a whole new interactive website (see www.solarstormwatch.com for an example).

Each project is supported by the Zooniverse Application Programming Interface (API). It was designed primarily as a tool for serving up a large collection of assets (e.g. images/ video) to an interface and collecting back user-generated interactions of these assets which could be as simple as a classification but as complicated as the results of a simulation tuned to fit the appearance of the asset. It was designed from the start be language-agnostic, that is, provided that the client software or interface can create and receive HTTP messages in XML or JSON formats it will be able to interact fully with it.

Testing, launch and data reduction

In any case once the main interface has been developed, it will be necessary to carry out a beta test to ensure that the data is produced is of an acceptable standard. Earlier stages of testing may involve local volunteers or with focus groups, but typically the beta phase will be open to existing Zooniverse volunteers. Where the task is simple enough, a tutorial can be optional (as in Galaxy Zoo) but where a higher level of understanding is necessary it may, following the results of the beta, be found necessary to make the tutorial compulsory (as in Solar StormWatch).

Following the successful operation of the project, a certain amount of data verification and reduction is necessary. This is normally also the responsibility of the project's science team. At minimum, this will involve comparison of the data collected from citizen scientists with professional or machine learning results, but may of course be more involved, assigning users weights depending on performance or measuring and accounting for subtle biases.

As the aim, and our obligation to those who volunteer their time, is to generate as much science as possible from their efforts, we would either expect the science team to have the resources to generate a significant collection of scientific products, in terms of publications, high-level public catalogues, etc., or to make the reduced dataset publicly available. The ideal scenario might be a short proprietary period, while the team performs their highest priority studies, followed by publication of the reduced dataset. We feel such a proprietary period is justified to reward the science team for their input in designing the project and as a quality control procedure before the full dataset is made public, allowing identified wrinkles to be ironed out prior to release to the community.

Beyond the primary interface

These steps – task identification, data set selection, beta testing and data reduction – are the bare minimum expected inputs from the science team for a citizen science project. In addition, though, we have found that projects are most successful when efforts are made to communicate results back to the volunteers, and we would expect scientists to blog and to be available to answer selected questions drawn from the forum. Secondly, much of the power of the method lies in the ability of advanced volunteers to follow up themselves on interesting or unusual data. This necessitates the provision of access to as much additional data as possible. This access can be kept separate from the main task to avoid contaminating classification, but is a powerful way

of making sure that a data set is properly mined. There is, of course, a huge scope for formal education too, and a tool to allow people to share resources has also been developed.

Now what?

The level of funding required to develop the interface depends on the chosen design, specifically on much can be reused from previous projects. As the current rate-limiting step is the amount of developer time available, projects which essentially reuse existing features (or which have enough funding to pay for their own development needs) can proceed quickly, and anyone is welcome to get in touch with us to begin planning a project. In the next few months, we will be making a series of open calls for projects which require substantial support from the Zooniverse team and anyone with ideas is welcome to apply.

Chris Lintott & the Citizen Science Alliance
chris@zooniverse.org
www.citizensciencealliance.org

CSA Questionnaire

What if you had hundreds of thousands of people ready and waiting to help analyse your data - would you collect it in a whole new way and tackle previously impossible scientific questions? Or perhaps you already have interesting research questions and a mountain of data, but not enough time and resources to make the most of it? If there is a critical role for human interaction in the analysis process then the Citizen Science Alliance (CSA) may be able to help by coordinating large-scale public involvement.

The purpose of this questionnaire is to assist with evaluating the suitability of potential CSA projects. Considering the answers to these questions should help your team identify which project ideas are feasible and determine which might benefit most from the Citizen Science approach. Combining this information with your science priorities will enable you to select ideas to propose as future CSA projects. We will review the information provided for your highest priority project proposals to check their suitability and determine the resources required to successfully execute the project. At this stage only very short answers and rough estimates are required ("no idea" is a valid answer to many of the questions). The details of the projects will be refined once they are selected for development.

Please bear in mind that the interests, abilities and commitment of the Citizen Scientist population are varied, so that even quite complicated or apparently uninteresting tasks may be feasible. Ambitious ideas are highly encouraged, particularly if they have a strong scientific need. A wide range of tasks are possible, some of which we already have the capability for from developing other projects, and others that we could work with you to develop. An impression that the data will not be sufficiently appealing to the public should not be a concern, as there are various ways in which the presentation of data and the user experience can likely be enhanced to attract sufficient participants, particularly if there is a compelling science case.

Below we use the term "raw data" to mean the data as provided by you or an external source (which will usually have been through some form of processing), that forms the basis of a task in a CSA Project, possibly after further processing by the CSA. We refer to data obtained from the tasks of the CSA Project, for example the individual clicks, question answers, measurements, etc. as "collected data".

Questions:

Provide a brief description of the science addressed by the proposed project.

What would the minimum achievements for success be, and what extra might you hope to get?

Describe the nature of the data that would be used in the proposed project. (Please provide some sample data if easily available.)

What specific tasks do you envisage citizen science participants performing with this data in order to address your science requirements?

What automatic processing routines exist which try to do the same job as the envisaged tasks? Why can't they be used instead of humans?

What makes your project particularly appealing to a non-specialist audience?

Are there any 'easy targets', such as existing interest groups, online communities or clubs?

What training will be required before a citizen-science participant can start performing tasks? Would it be possible for volunteers to 'jump straight in', with a suitable interface, or is some training essential?

How much time do you estimate each task (or group of tasks) will take?

If possible, estimate the minimum number of times a task must be performed on a given element of data to be useful for science (assuming all tasks are performed by competent citizen scientists)? (Once might be enough for clear cut tasks, more times could be required for fuzzier tasks, lots of times may be necessary if accurate estimates of uncertainties are desired.)

Very roughly, estimate the total number of tasks that must be completed before your research goals will be achievable.

Does the raw data currently exist, or when will it be available? Is it dependent upon securing further funding?

Is the raw data in any way proprietary and, if so, what measures will be necessary before it may be displayed to the public (e.g. concealing coordinates, blurring faces)?

What further processing is required before the raw data can be used as an input for the envisaged tasks (in a 'web-ready' format, e.g. jpeg)? (Will you be able to do the required processing, or do you need assistance?)

Describe the estimated volume of the raw data (e.g., size of each image, number of images, total disk storage required).

How many individual elements of data are there in total (e.g., 1 million images to be classified)?

What is the envisaged final form of the collected data for science use?

Will the final collected dataset itself be of legacy value, and worth making publicly available? (Will its use by others require additional proprietary information? Are you willing to make such information public along with the collected dataset to maximise its value to the wider community?)

Have you considered how you might process the database of individual task results into the envisaged final form?

Do you already have a science team to analyse the collected data, particularly during any proprietary period? (If not, do you envisage forming such a science team?)

What funding do you have already, anticipate applying for, and require in order to perform the analysis you desire on the collected data?

What are the anticipated minimum public outputs from this project (e.g., catalogues, articles, etc.), and their timescales once sufficient data is collected?

What is the expected impact of the work resulting from this proposed project on the relevant research community?

Are their potential extensions to the project that you already have in mind?

Are there members of your science team who are willing to write blog posts regarding the results coming from the primary task?

Are there members of your science team who are willing to monitor forum threads related to the science topic(s)?

Does this project tie in with any public engagement activities you are involved with?