Zooniverse: Observing the World's Largest Citizen Science Platform

Robert Simpson
Department of Physics
University of Oxford
United Kingdom
robert.simpson@astro.ox.ac.uk

Kevin R. Page
Oxford e-Research Centre
University of Oxford
United Kingdom
kevin.page@oerc.ox.ac.uk

David De Roure
Oxford e-Research Centre
University of Oxford
United Kingdom
david.deroure@oerc.ox.ac.uk

ABSTRACT

This paper introduces the Zooniverse citizen science project and software framework, outlining its structure from an observatory perspective: both as an observable web-based system in itself, and as an example of a platform iteratively developed according to real-world deployment and used at scale. We include details of the technical architecture of Zooniverse, including the mechanisms for data gathering across the Zooniverse operation, access, and analysis. We consider the lessons that can be drawn from the experience of designing and running Zooniverse, and how this might inform development of other web observatories.

Categories and Subject Descriptors

D.2.11 [**SOFTWARE ENGINEERING**]: Software Architectures; H.3.4 [**INFORMATION STORAGE AND RETRIEVAL**]: Systems and Software; J.2 [**PHYSICAL SCIENCES AND ENGINEERING**]: astronomy

Keywords

web observatories; citizen science; crowdsourcing

1. INTRODUCTION

The development of new techniques to handle increasing quantities of digital data has led to the creation of online platforms to distribute data analysis: a type of citizen science. The crowdsourcing of hundreds of thousands of people in the scientific process has proven to be a technique capable of making a valuable contribution to this problem. The human combination of intuition with abilities in pattern recognition and analysis can, for a wide variety of scientific tasks, far outstrip the performance of even sophisticated automatic systems.

Zooniverse projects invite the public to participate in genuine data analysis at a scale that researchers cannot accomplish on their own (or even in sizeable groups). Research data is shown to users in the form of images, video and audio via one of the Zooniverse websites. Volunteers are shown how to perform that required analysis via a simple guide or tutorial such that they can then identify, classify, mark, and label them as researchers would do.

The first Zooniverse project, Galaxy Zoo [4, 3], launched in July 2007 and successfully engaged 165,000 volunteers in the morphological classification of images of galaxies. The early success of this first project led the team behind it to explore new research domains and types of task and user interface.

Zooniverse citizen science projects have resulted in the classification of more than a million galaxies, the discovery of nearly a hundred exoplanet candidates, the recovery of lost fragments of ancient poetry, and the classification of more than 18,000 thousand wildebeest in images from motion-sensitive cameras in the Serengeti. The combined effort of hundreds of thousands of volunteers adds up to more than 50 years of non-stop effort each year on the Zooniverse platform alone, but without further fundamental development it will not be able to cope with future data volumes.

At the beginning of 2014 there were 20 live Zooniverse projects, covering topics including space, nature, biology, medicine, climate science and the humanities[1]. Over 900,000 volunteers were registered to the site (915,000 accounts had been created); although it should be noted that you can participate on many projects without creating an account so this number is a lower-limit on the true volume of participants. Seven projects launched in 2013 and two more in January 2014; a further three will be active by March 2014.

At the time of writing, results from the various Zooniverse projects have been included in more than 250 scholarly articles[2], with 59 dedicated to Zooniverse findings; 39 publications have resulted from Galaxy Zoo alone. Examples of particular impact are the data release of the original Galaxy Zoo project, including all raw and reduce data [4]; the analysis of crowdsourced freeform annotation of star-formation on our Galaxy resulting in an order-of-magnitude detection rate for infrared bubbles [8]; and the discovery of unusual exoplanetary systems detected by volunteers and missed by automated routines [2, 7].

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579215.

[1]A complete list of Zooniverse projects can be found at `http://zooniverse.org/`

[2]A complete list of publications across all the Zooniverse projects can be found at `http://zooniverse.org/publications`

2. ZOONIVERSE DOMAIN MODEL

The Zooniverse domain model pervades the platform and is used and adhered to by both the back-end storage facilities and in-browser Javascript - the platform architecture, in effect, is a mechanism for scaling and distributing records according to the model.

The domain model is relatively simple, focussing on the parts most critical to Zooniverse operations and the instantiation of projects, rather than the detailed domain facets unique to each project:

User. People are core to the Zooniverse. A User is a cross-platform, i.e. potentially multi-project, entity. It represents a person/account with a username, an email address, and information about the projects in which they have participated.

Subject. These are the things that users classify, annotate or transcribe. In Old Weather it's a scanned image of a ship logbook, in Planet Hunters [2] it's a light curve. Subjects are what is shown to users on the site for classification. Subjects are associated with additional metadata, useful for a project's researchers, such as geographic location of origin, camera/telescope details, field-size, etc.

Workflow/Task. When a User is presented with a Subject on one of our projects we ask them to do a Task. These Tasks can be grouped together into a Workflow. Most projects just have a single Workflow. In Notes from Nature each step of the transcription is a separate Task, in Galaxy Zoo, each step of the decision tree is a Task too.

Classification. Classifications are the core Zooniverse unit of human effort. A Classification represents what a person saw and what they said about it.

Group. Subjects can be grouped together for higher level purposes. A Group can represent an astronomical survey in Galaxy Zoo (such as the Hubble CANDELS survey) or a Ship in Old Weather. These groups often require slightly different display rules, for analysis pipelines.

Project. This is simply the citizen science project itself, with which Subjects, Classifications and Groups can be associated. Since the Zooniverse platform hosts many varied projects it is useful to distinguish between them and the other elements associated with each.

These entities are common to all projects implemented on the Zooniverse platform, and so provide both a common information substrate upon which all Zooniverse projects are built and extended. They also give an indication of the properties which have provided most cross-project utility during software development over the last 6 years.

3. ZOONIVERSE ARCHITECTURE

The Zooniverse is heavily reliant on Amazon Web Services (AWS), particularly Elastic Compute Cloud (EC2) virtual private servers and Simple Storage Service (S3) data storage. AWS is the most cost-effective solution for the dynamic needs of Zooniverse's infrastructure and in 5 years of operation on the platform there have been less than 5 large-scale problems – while placing all your eggs in one basket creates a single point of failure it also reduces dependence on multiple third-party services.

Originally each new project was operated as its own Rails application on its own EC2 instance. This soon became expensive and difficult to maintain. We have now moved to using a large-scale, central datastore and API which communicates with projects that are thick-client pure-JavaScript applications (figure 1). This provides an interesting insight into the commonalities that can be observed across all the projects Zooniverse supports.

All Subjects (per the domain model) are stored on S3, as are the thick-client applications for each project. This creates a fast user-experience (once an app has loaded) and consolidates load in a central API provider which can be scaled to adjust to the level of traffic in the system – load typically being increased by an underlying increase in Users, Subjects, or Classifications. Uniformity of a core API between projects enables simple addition of extra cloud resources to meet demand, while still allowing a custom user interface design for each project which is useful in tailoring projects to a community's needs, e.g. a planet-hunting website needs a very different look-and-feel to an animal-spotting website.

The Zooniverse's central, large, Mongo datastore is managed by a custom Rails application called 'Ouroboros'. Ouroboros knows where all a project's Subjects are kept and which User has seen which subjects. It intelligently assigns and delivers Subjects for Classification to Users via a simple JSON API accessed from by the JavaScript running in the User's web browser. It also receives classifications and stores them for later analysis, delivering data to various science teams as needed and on-request. At time of writing there have been more than 500 million classifications across Zooniverse project, and more than a million unique participants.

Ouroboros anonymises data before delivery to researchers, though some personal data is stored, such as IP address and browser-header, as these can be very useful in data reduction and clustering of results.

4. DATA ACCESS AND API

The Zooniverse platform in 2014 is built upon the Ouroboros API, which operates (via JSON) the backend Ouroboros Rails application. This API is used internally for virtually all create, read, update and destroy actions. The project web-apps utilise a custom JavaScript library (called the Zooniverse library), and a templated, MVC JavaScript tool to enable the quick, easy creation of new projects in a simplified, wireframe form. Design and user-experience customisations are then implemented. A clone of the main Ouroboros API exists to allow for prototype development and testing.

The Zooniverse JavaScript library abstracts all aspects of standard Zooniverse behaviour by replicating the main domain model objects as JavaScript classes. The library handles CRUD actions by communicating asynchronously with the main API, taking advantage of HTML5's local storage capabilities to buffer Subjects and Classifications between the front- and back-ends of the system.

The Ouroboros API exposes HTTP resources with a JSON representation for data instances conforming to the data

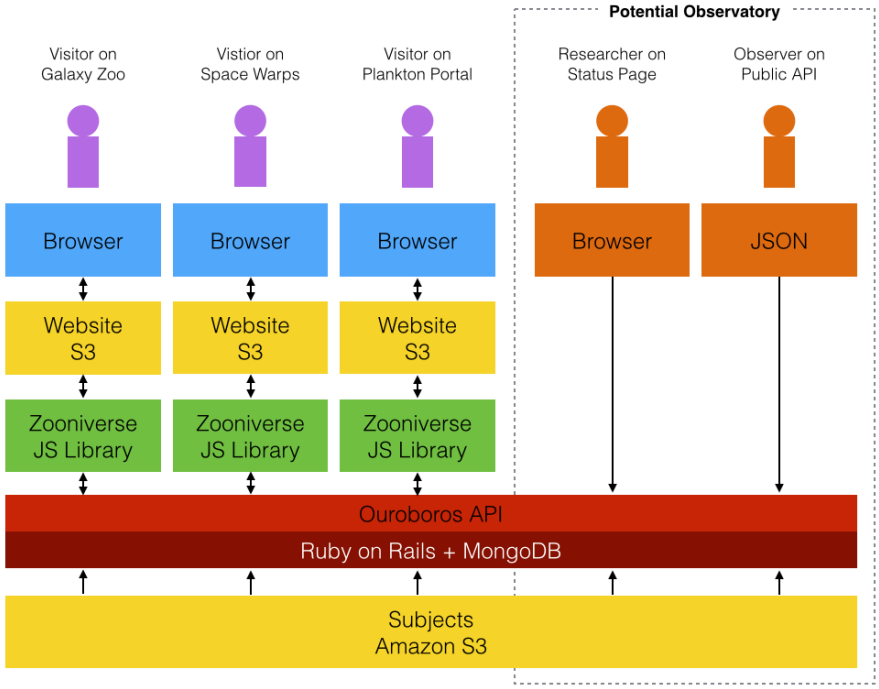

Figure 1: The Zooniverse Architecture

model. For example a project URI[3] represents the Milky Way project and serves JSON of the form:

```
{ "activated_subjects_at" : "2014-01-27T22:00:13Z",
  "bubble_count" : 150043,
  "classification_count" : 557657,
  "cluster_count" : 119576,
  "complete_count" : 50999,
  "created_at" : "2013-12-10T17:56:45Z",
  "display_name" : "The Milky Way Project",
  "ego_count" : 70334,
  "galaxy_count" : 38021,
  "groups" : {
      "523ca1a03ae74053b9000003" :
                { "classification_count" : 179924 },
      "523ca1a03ae74053b9000004" :
                { "classification_count" : 377733 }
  },
  "id" : "523ca1a03ae74053b9000001",
  "name" : "milky_way",
  "site_prefix" : "MW",
  "updated_at" : "2013-12-10T17:56:45Z",
  "user_count" : 3367,
  "workflow_name" : "milky_way",
  "zooniverse_id" : 7
}
```

While a Subject resource[4] returns a JSON representation of a subject within the Milky Way project:

```
{
  "activated_at" : "2013-12-10T19:15:49Z",
  "classification_count" : 9,
  "coords" : [ 2.1093000000000002, 28.971800000000002 ],
  "created_at" : "2013-12-10T18:11:30Z",
  "follower_ids" : [ "505026fb0454e27ad8001e47" ],
  "group" : { "_id" : "523ca1a03ae74053b9000003",
      "name" : "glimpse3d",
      "zooniverse_id" : "GMW0000001"     },
  "group_id" : "523ca1a03ae74053b9000003",
  "id" : "529e522e4d696359f4e23100",
  "location" : { "standard" :
    "http://www.milkywayproject.org/subjects/529e4.jpg" },
  "metadata" : { "file_name" :
                    "GLM_028.9718+02.1093_mosaic_I124.jpg",
      "markings" : { "blank_count" : 5,
          "bubble_count" : 0,
          "cluster_count" : 1,
          "ego_count" : 1,
          "galaxy_count" : 0
        },
      "size" : "0.1500x0.0750"
    },
  "project_id" : "523ca1a03ae74053b9000001",
  "random" : 0.3657572225983644,
  "state" : "complete",
  "updated_at" : "2014-01-10T17:34:52Z",
  "workflow_ids" : [ "523ca1a03ae74053b9000002" ],
  "zooniverse_id" : "AMW00009ur"
}
```

[3] https://api.zooniverse.org/projects/milky_way/
[4] https://api.zooniverse.org/projects/milky_way/subjects/AMW00009ur

Public read access is available for science teams and Zooniverse developers, and the status of any project can be accessed by all members of the team (i.e. developers, researchers, educators). The Ouroboros API provides information of recent participation and classification numbers (hourly, daily, weekly) as well as an overview of the completion of the dataset, e.g. stating that 42% of the subjects have been fully-classified within the defined rules assigned to that project (figure 2). Google Analytics is also used to track real-time web usage and details of who uses the site, how they use it and where they were referred from.

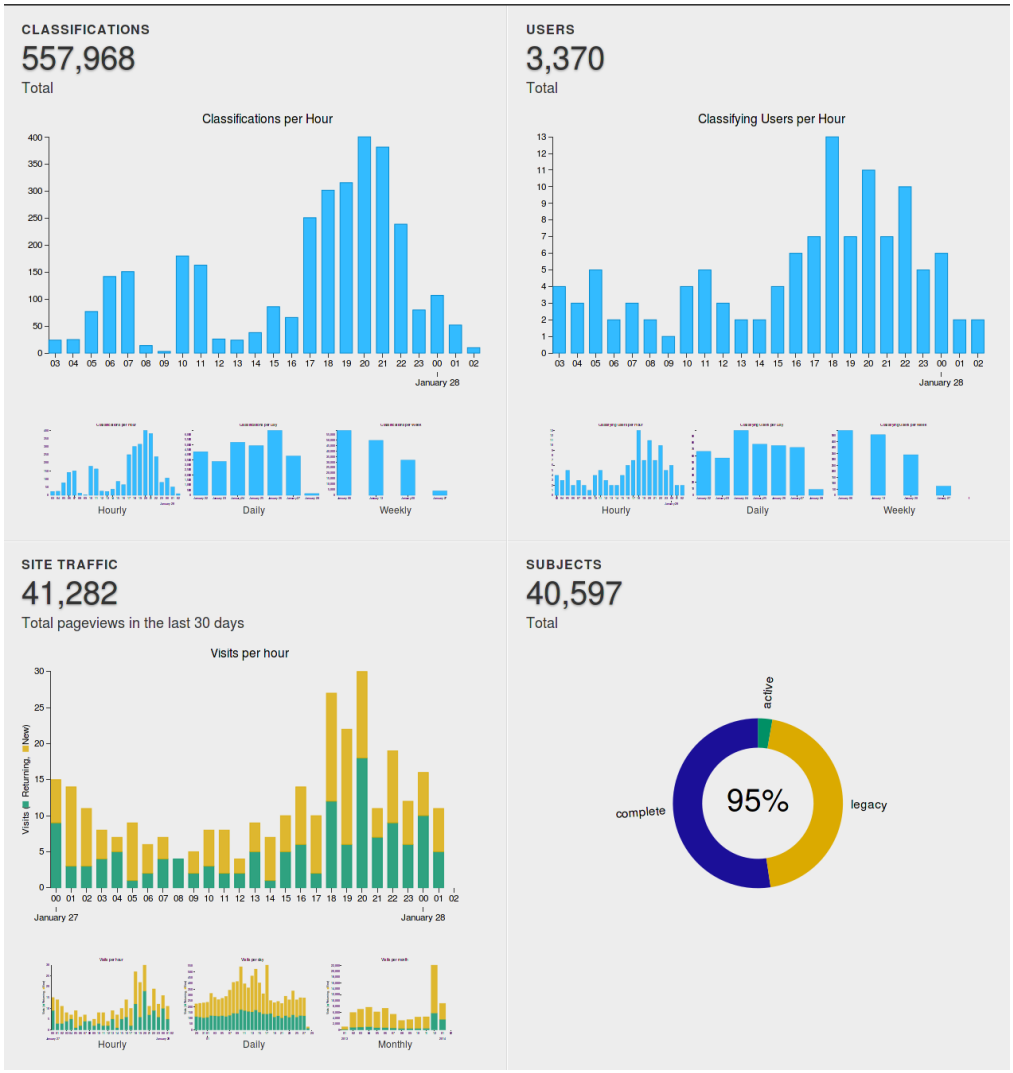

Figure 2: Example of Zooniverse project status (for the Milky Way project)

5. DESIGN LESSONS

The creation of engaging user experiences is essential to getting the best from volunteers online. While Ouroboros generalises the common tasks between projects, the user experience must frequently be different for a specific project to maximise user interest and participation. Hence the Zooniverse has evolved into a flexible platform for web developers, which has led to a wide range of high-quality Zooniverse projects. Testing with real users has become vital to the project development process; however the most dedicated volunteers from amongst the user base can sometimes be those who are also the most unwilling to accept change. We have come to recognise that predicting success is difficult, but that a key factor in the 'stickiness' of our projects is the 'cool' factor people perceive in them, such that 'discover your own planet' is cool but 'watch genetically-modified worms lay eggs' may not be.

The Zooniverse domain model used today is informed by six years of project-building. Since Old Weather in 2010, projects have been built for diverse research domains and so we need a core Zooniverse domain model that's flexible enough to support vastly different projects, but not overly simplified so that we are prevented from building rich user experiences.

Our domain model has also been heavily influenced by the patterns that have emerged working with science teams. Appropriate levels generalisation to abstract complexity away from individual project development have been key, albeit achieved by iterative evaluation and modification. In the early years we recorded each step of a User's Classification of a Subject as a distinct descriptive item called an Annotation. The science teams did not use them and they were irrelevant to realtime operations. The vast majority of Zooniverse projects to date collect large numbers of Classifications that are write once, read much, much later.

Embracing this fact means that the platform can be less concerned about exactly what is stored at any given time and instead focus on ensuring that data is reliably and repeatedly stored – that the platform has high availability and can operate and scale to meet peak demand, using an architecture and domain model derived from these operational requirements. The data structures stored are those that are

best for researchers to actually use, some time after collection when further analysis takes place.

Thus, in the most-recent projects much of the same information is being stored, but without needing to force it to conform to a specific part of a specialist model. Now all information about the components of a classification are simply metadata records within the classification itself. This is another reason to use a NoSQL datastore such as MongoDB: the flexibility exists to store arbitrary metadata about classifications from diverse projects in one database.

6. ZOONIVERSE WITHIN THE WEB OBSERVATORY ECOSYSTEM (RELATED AND FUTURE WORK)

There are broadly two contexts through which we can consider Zooniverse as an element of the emerging Web Observatory ecosystem and envisage how this might be developed in the future: (i) as an observatory in its own right, where we might apply Web Observatory tools to further the study of Zooniverse, and (ii) as part of a larger Observatory where aspects of Zooniverse are investigated in conjunction with other observation sources, raising issues of interoperability and data alignment.

Considering an observatory platform such as Truthy [5] we can see numerous common functional goals and concerns shared by Zooniverse: reliability, reproducability, topic filtering, and visualization. The engineering in the two platforms, however, is directed towards significantly different causes: in Truthy a significant proportion of observatory capability is in clustering tweets (e.g. into memes) which can be explored through the visualisations and which are exposed via the Truthy API; in Zooniverse such groupings (and "data-driven views") are already explicit in the Ouroboros data model and API, with engineering focused on efficiently and reliably gathering data in real time from users. The difference, then, comes from the scope and source of data collection: pre-determined specific information collected directly from humans for Zooniverse, versus completely generic emergent information collected from an intermediate communications infrastructure (Twitter) for Truthy. These entirely reasonably generate divergent engineering requirements.

On the other hand we can see that, compared to the Ouroboros status view (figure 2), Truthy provides a much more sophisticated view over computed statistics, interactive diffusion networks and exploration interfaces that might enable a deeper and more rigorous examination of Zooniverse behaviour – after all it is highly desirable to gain a greater understanding of which projects are more successful than others and why (beyond 'watching genetically-modified worms lay eggs' not being cool!). An immediate possibility and question, then, is whether platforms such as Truthy can be adapted to ingest data from Ouroboros rather than only from Twitter tweets.

In Zooniverse there is already a strong data model and a clear boundary in purpose (citizen science) that in many ways constrains the potential topics that might be investigated when Zooniverse is considered as an isolated observatory. Of course Zooniverse does not exist in a vacuum, and there are social and computational interactions worthy of study when we consider multi-part socio-technical interactions as trajectories [6]. For example, is it possible to study and build more effective incentive mechanisms purely based upon analysis of Ouroboros derived information, or must this data be combined with other sources of quantitative and/or qualitative investigation to gain insight?

Thus it would seem likely that not all questions can be answered in isolation, and so others will involve investigation through an observatory that combines multiple data sources from different systems and with different data models. For example, a hypothesis might be that when a Zooniverse project receives broadcast publicity (e.g. through national television or radio) this attention is amplified through social media and leads to an increase in volunteers drawn from the general public. i.e. the object of observation comes to specifically include the engagement of volunteers. Can this be quantified? How might an observatory bring together Zooniverse and Twitter data that documents the wider public visibility of an experiment? Through a simple temporal alignment with hyperlinks to the project in tweets? Given its social media observatory credentials could a platform like Truthy, or others, be adapted to this end? Can this be combined with a qualitative analysis of activity in the Zooniverse forums?

Brown et al. [1] identify potential features of observation, including those related to collaboration between parties and the communication required to exchange (and potentially standardise) data that is complex and comes from multiple sources. They identify a formal classification according to topics, which we have discussed above when comparing Truthy to Zooniverse, as one critical feature. Through our study of Zooniverse, we believe careful consideration must be given as to how such interoperability (e.g. of topic classification) is achieved; while Zooniverse does encapsulate individual topics, this is not directly exposed through the Ouroboros API since that has been designed to primarily meet operational, rather than observatory, needs. We do not believe this makes Zooniverse any less a Web Observatory, but certainly one whose capabilities – and the evolution of its design – should be considered and embraced when considering requirements for interoperability.

7. REFLECTIONS ON WEB OBSERVATORIES

Zooniverse provides a capable and interesting example of a web observatory with a clearly scoped remit, and of great potential for further investigation when combined with other observatory sources – not to mention the significant challenges in unifying such content.

The evolution of Zooniverse, as described in this paper, demonstrates the design decisions a platform must take to become a reliable service that operates at scale. Earlier iterations of the system software gave more focus to the nuances and data of individual projects – which are more immediately suitable for exposing domain information to other observers and observatories, where such breakdown of behaviour is desirable. The platform as it stands today has a more streamlined design driven by operational requirements and the need to quickly start new projects and scale to meet rising (and falling) demand. An interoperable observatory layer would need to complement and extend Ouroboros, not replace it.

Typically discussion around building larger Web Observatories has concentrated on the interoperability layer; of

exposing common descriptions and properties between compatible observatories. These are crucial, of course, but Zooniverse provides a counterpoint – that we must also consider how to scale, and how a scalable design will influence our data models and APIs. Engineering for scale is often a secondary motivation to Web Observatory interoperability[5], but as Zooniverse illustrates it can drive a radical reversal in the data elements an API offers, which in turn frames how the platform can be presented as an interoperable Observatory. Finally, we note that although the ever increasing size of Zooniverse is impressive, it is but a fraction of the larger analyses that could be performed over an aggregated network of Web Observatories – size matters.

Acknowledgements

The authors would like to thank the Zooniverse team for their support writing this paper and more generally in their work during development of the platform described herein. This work is supported under SOCIAM: The Theory and Practice of Social Machines, a programme funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1, and a collaboration between the Universities of Edinburgh, Oxford, and Southampton; and also by funding from the European Community's Seventh Framework Programme (FP7/2007-2013) Smart Society project (FP7-600854).

8. REFERENCES

[1] I. Brown, W. Hall, and L. Harris. Design and prototyping of a social media observatory. In *Proceedings of the 1st International Web Observatory Workshop, 22nd International Conference on World Wide Web companion*, pages 1317–1320. International World Wide Web Conference, 2013.

[2] D. A. Fischer, M. E. Schwamb, K. Schawinski, C. Lintott, J. Brewer, M. Giguere, S. Lynn, M. Parrish, T. Sartori, R. Simpson, et al. Planet hunters: the first two planet candidates identified by the public using the kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, 419(4):2900–2911, 2012.

[3] L. Fortson, K. Masters, R. Nichol, K. D. Borne, E. M Edmondson, C. Lintott, J. Raddick, K. Schawinski, and J. Wallin. Galaxy zoo: Morphological classification and citizen science. *Advances in Machine Learning and Data Mining for Astronomy*, pages 213–236, 2012.

[4] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178, 2011.

[5] K. McKelvey and F. Menczer. Design and prototyping of a social media observatory. In *Proceedings of the 1st International Web Observatory Workshop, 22nd International Conference on World Wide Web companion*, pages 1351–1358. International World Wide Web Conference, 2013.

[6] K. Page and D. De Roure. Trajectories through social machines. In *Proceedings of the 1st International Workshop on Building Web Observatories at ACM Web Science*.

[7] M. E. Schwamb, J. A. Orosz, J. A. Carter, W. F. Welsh, D. A. Fischer, G. Torres, et al. Planet hunters: A transiting circumbinary planet in a quadruple star system. *The Astrophysical Journal*, 768(2):127, 2013.

[8] R. J. Simpson, M. S. Povich, S. Kendrew, C. J. Lintott, E. Bressert, K. Arvidsson, et al. The milky way project first data release: a bubblier galactic disc. *Monthly Notices of the Royal Astronomical Society*, 424(4):2442–2460, 2012.

[5]We note that Truthy, for all the features and benefits referenced above does "not scale to [this] expanded collection" [5].