Community-as-a-Service: Data Validation in Citizen Science

Yurong He and Andrea Wiggins

College of Information Studies, University of Maryland
College Park, MD, USA

{yrhe,wiggins}@umd.edu

Abstract. Currently, most citizen science projects that adopt a crowdsourcing model focus primarily on collecting and analyzing data. As yet, few of them leverage community interactions for effective data validation yet, which would have significant impact on improving the quality of the increasing volume of citizen science data. In this paper, we introduce an exploratory pilot study focused on understanding how an established online community can be leveraged to create a "*community as a service*" structure to support collaborative citizen science data validation.

Keywords: Community as a service, crowdsourcing, data quality, citizen science, data validation

1 Introduction

Advances in information technologies have dramatically increased our capabilities for collecting, processing, and accessing large volumes of data, leading to new challenges in evaluating and assuring data quality. Although it would be ideal if all data quality assessment and improvement tasks could be automatically accomplished by machines, nearly every data set includes records that require direct human attention in some fashion due to the current limits of computing capabilities. With increasing volumes of data that require manual processing or verification, one potential solution is to seek help from "the crowd". In recent years, crowdsourcing—a mode of outsourcing that involves a large number of (typically) unknown contributors [1]—has been applied to validating big data created by both humans (e.g., user-generated content), such as volunteered geographic information (VGI) [2] and machines, such as satellite data [3]. Data validation is very important from the data use and repurposing perspective, especially if the data are collected by non-experts for solving scientific problem or making policies and laws (e.g., citizen science).

The members in the crowd are usually independent from and relatively unknown to each other, and are asked to help accomplish predefined highly simplified, repetitive, and straightforward tasks that require little domain expertise [4]. However, some data validation tasks are difficult and complex, for example, confirming the likelihood of a

rare species being observed at a particular time and place. The skills and knowledge needed to evaluate these data are a less obvious fit to crowdsourcing strategies, but as data sources continue to grow, accumulating a large number of small contributions from diverse contributors remains one of the most sustainable approaches to validating data that require informed judgment.

We therefore introduce the concept of a "*community as a service*" model to tackle the problem of difficult data validation tasks. Communities consist of people, like the crowd, but the members of a community are more interdependent and usually come together around shared interests, social identities, and norms, which typically lead to higher levels of commitment to the tasks and stronger ties with each other than in most crowd contexts [4]. Community members are willing to put more time and energy into more difficult tasks if the tasks fit their individual interests as well as the community interests. Crowds and communities are often framed as two ends of a spectrum. In this study, we focus on a group that is located closer to the community end of the spectrum, as observed through evidence of socialization, offline interactions, and coordinated collective efforts. Based on a preliminary analysis of this community's activities, we propose that seeking help from "the community" may prove useful when knowledge-based manual data processing or verification is needed.

Citizen science offers a number of examples of crowd-created big data, such as the precipitation network data from the Community Collaboratory on Rain, Hail, and Snow, and the bird observation data from the eBird project (ebird.org). If citizen science communities can in fact leverage community interactions for effective data validation, this could become a viable service offering for sustaining citizen science tools in a strategy similar to other technology support services, with community-based data validation as the service rendered in a "*community as a service*" model. Naturally, community-as-a-service offerings would have to generate net benefit to the community or involve community members in governance to avoid concerns around exploitation; however, as part of a sustainability strategy, this may be feasible for some citizen science projects.

We therefore investigated whether an established online community and its platform can provide robust data validation services for citizen science projects that need to collect and verify biodiversity observation data. This exploratory pilot study focused on understanding how such an online community can be leveraged to create a community-as-a-service structure to support data validation by examining the existing data validation interactions on a social computing platform, iNaturalist.

2 Related Work

Citizen science is a type of research collaboration that often resembles crowdsourcing, in which volunteers contribute their efforts to help advance scientific research [5]. The general expectation is that data from non-expert volunteers are more prone to error, especially when the tasks of producing data include difficult or complex steps, such as biological specimen identification [6, 7]. The way organism identifications are verified in traditional biology is via collecting physical vouchers in the field and ana-

lyzing the vouchers in the lab (e.g., extracting DNA from an insect leg). Instead, in most citizen science projects, the physical vouchers are replaced by digital vouchers (e.g., photo, audio) that are uploaded to and verified in online environments [8]. This is a trade-off for the ability to collect large amounts of data over wider spans of space and time. Expert review is considered an indispensable step of ensuring the quality of this type of data [9].

Currently, the most effective way of verifying organism observation data online is combining machine and human intelligence [10, 11]. For example the eBird project, in which volunteers contribute wild bird observation data, developed a two-step data quality improvement approach that includes automated filters and a network of regional experts [10]. The machine automatically identifies questionable data based on comparison to prior data, asks observers to confirm whether the data is legitimate, and if yes, a pre-identified regional expert is asked to help review the data; this individual then communicate with the observer to get more supporting information in order to verify the observation [10].

However, a drawback of an expert-driven approach is that the number of trained, professional experts is still too small to verify all such data that needs human verification [9], suggesting that combining experts' efforts with those of a larger number of volunteers may be valuable for verifying observation data. Most citizen science projects build and sustain communities of volunteers to collect and analyze science data. The members of these communities include both professionals and amateurs, with some shared interests in a certain domain or specific topic, who interact with each other to varying degrees in both online and offline environments [9]. But few of these projects have turned to their community of volunteers for collaborative data validation, which could have significant impact on improving the quality of the increasing volume of citizen science data. This is promising because most citizen science projects have too little baseline data for algorithmic identification of outliers, are constrained by information systems that were not designed to support distributed review or data validation, and do not have the resources to support expert review at increasing scales.

3 Study site

iNaturalist (http://www.inaturalist.org/) is a social network site for professional and amateur naturalists, with additional functionality that supports independent, unaffiliated citizen science projects with data collection and management of biodiversity observation data. As of June 2015, the iNaturalist community has over 70,000 users who had contributed over 1,550,000 observations. Anyone can record any organism they observe in nature, meet other nature lovers, and learn about the natural world on iNaturalist. Similarly, any citizen science project focused on observing living organisms can create a project page and invite community members to contribute data and assist in data validation.

4 Methods

For the pilot study, we adopted a multi-level case study methodology [12]. At the social computing platform level, we examined the iNaturalist community dynamics. At the level of a project hosted by iNaturalist, we investigated the community data validation interactions on a citizen science project, Biocube (http://qrius.si.edu/biocube), which collects data on biodiversity and uses iNaturalist for data management.

We first downloaded the dataset from iNaturalist Export Observation Page (http://goo.gl/EiHy0a) including 972 organism records observed on January 24th, 2015 in the United States. Each record contained 61 attributes reflected by the interface on the observation record page. We then chose 39 observations and the contents on each observation record page from the Biocube event on January 24th, 2015 as the focal data for initial exploratory analyses. A total of 25 participants included ten science teachers and three educators who work at nature centers. The remaining participants were facilitators, including two social scientists, two professional photographers, three educators, and five biologists. After collecting data in the field, participants were asked to submit data on iNaturalist. We observed the entire event and monitored the processes on iNaturalist through which the records were improved over the next five months. Data used for analyses reported here were directly downloaded or manually collected from public-facing pages.

5 Results

Within the larger data set of 972 records submitted in the US on the date we selected, there were 39 records uploaded to iNaturalist for the Biocube event. Each record represents a specific organism, submitted with an identification (name or ID) of the organism, when and where it was observed, at least one photo, and a label associating it with the Biocube project. If the participant was uncertain about the organism ID, an "ID please" flag could be added to the record to request assistance from other community members. Table 1 shows the count of records in the top level taxonomic categories for the Biocube event on iNaturalist.

Table 1. The number of records in each top level taxonomic category. "Something" represents an observation for which there was no organism ID entered.

Taxon	Actinopterygii	Animalia	Arachnida	Insecta	Mollusca	Plantae	Something
Count	3	14	1	5	5	10	1

After each record was created, iNaturalist community members could help verify data in three ways: (1) agreeing with the organism IDs provided by the observer and/or other iNaturalist community members; (2) suggesting a new organism ID, usually a more specific label within the taxonomic hierarchy; and (3) leaving com-

ments. The basic information on each record and these data validation interactions are displayed like a threaded "history" on the record page, automatically documenting data provenance (e.g., who agreed with or suggested what organism ID, who left which comments, etc.)

After five months, community interactions on the focal records included: 17 records had no ID agreements, 13 records had one agreement, 6 records had two agreements, and 3 records had three agreements. 11 records had organism IDs suggested by community members. In some cases (N=9), these interactions helped refine the organism ID to a more specific taxonomic level. The smallest improvement of this type was from order to suborder, which is one level of change in the iNaturalist taxonomic hierarchy, and the largest improvement was from kingdom to species, a change of 18 levels; the average improvement was about nine levels, which represents a substantive improvement in ID specificity. For one record, organism ID became more general and less specific (genus to subfamily) because it was hard to identify the species with the available information, so the more general name better reflected the certainty of the ID.

There were four records with at least one comment from a community member. Comments asked for supporting details, like features not visible in the photos (e.g., whether a sea worm's body had sections) or their expected geographic distribution. In addition, our examination of records indicated that users often showed taxon-specific preferences for record validation (e.g., a highly skilled amateur who specializes in mollusk identification).

For this small sample, approximately 58% of the data were improved overall, where improvement means agreement by another user or refinement of the organism identification. At a minimum, the data were verified by at least one other individual. Almost 24% of the records were verified by two or more individuals, so in future work, we plan to explore how many agreements by community members would be considered adequate to replace expert validation, which has limited scalability. We also note that improvement to these records included not just confirmation or disconfirmation, the usual goal of data review in citizen science, but also refinement of the data that adds information content in the form of more specific organism IDs. This highlights the need for additional work to measure the added value from multiple types of validation, as creating a viable community-as-a-service offering would depend on such information.

6 Conclusion and Next Steps

In this paper, we described our initial observations of how the community validated and improved the data for a citizen science project that adopted iNaturalist as a distributed data management platform. Our next steps include:

- Increasing the number of data points (i.e., longer time periods and multiple projects) for in-depth analysis;
- Case-based comparisons of how different citizen science projects use iNaturalist and if variations in project-specific norms impact community data validation;

- Investigating how to measure and report value-added improvements to data;
- Using social network analysis to investigate the influence of community social structure on participation in general and data validation activities in specific; and
- Further developing the concept of "*community as a service*" by identifying evidence of similar fee-for-service (or service-for-service) arrangements in other contexts where volunteers provide the service as part of normal community activities.

Acknowledgements. This work was supported in part by NSF Grant 14-42668. We sincerely thank the hundreds of iNaturalist members who contributed the data used in these analyses.

References

1. Howe, J.: The rise of crowdsourcing. Wired, pp. 1-4. Dorsey Press (2006)
2. Fritz, S. et al.: Geo-Wiki. Org: The use of crowdsourcing to improve global land cover. Remote Sens, 1(3), 345-354 (2009)
3. Hansen, L.T., Ciciarelli, C.: Global forest watch-fires: Improving remote sensing through community engagement. In: 2015 AAAS Annual Meeting. (2015)
4. Haythornthwaite, C. Crowds and communities: Light and heavyweight models of peer production. In Proc. HICSS'09, pp. 1-10. IEEE. (2009)
5. Wiggins, A., Crowston K.: From conservation to crowdsourcing: A typology of citizen science. In Proc. HICSS'44, pp. 1-10. IEEE. (2011)
6. Conrad, C. C., Hilchey, K. G: A review of citizen science and community-based environmental monitoring: issues and opportunities. Environ Monit Assess. 176 (1-4), 273-291 (2011)
7. Royle, J. A.: Modeling abundance index data from anuran calling surveys. Conserv Biol. 18, 1378–1385 (2004)
8. Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S.: eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv, 142(10), 2282-2292. (2009)
9. Moran, S., Pantidi, N., Rodden, T., Chamberlain, A., Griffiths, C., Zilli, D., ... & Rogers, A.: Listening to the forest and its curators: lessons learnt from a bioacoustic smartphone application deployment. In Proc. CHI2014, pp. 2387-2396. ACM. (2014)
10. Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... & Kelling, S.: The eBird enterprise: an integrated approach to development and application of citizen science. Biol Conserv. 169, 31-40. (2014).
11. Bonter, D. N., & Cooper, C. B.: Data validation in citizen science: a case study from Project FeederWatch. Front Ecol Environ. 10(6), 305-307. (2012)
12. Yin, R. K.: Case study research: Design and methods. Sage publications. (2013)