Mechanisms for Data Quality and Validation in Citizen Science

Andrea Wiggins*, Greg Newman†, Robert D. Stevenson‡, and Kevin Crowston*
*School of Information Studies
Syracuse University, Syracuse, NY USA
Email: awiggins@syr.edu, crowston@syr.edu
†Natural Resource Ecology Laboratory
Colorado State University, Denver, CO USA
Email: newmang@nrel.colostate.edu
‡Department of Biology
University of Massachusetts at Boston, Boston, MA USA
Email: robert.stevenson@umb.edu

***Abstract*—Data quality is a primary concern for researchers employing a public participation in scientific research (PPSR) or "citizen science" approach. This mode of scientific collaboration relies on contributions from a large, often unknown population of volunteers with variable expertise. In a survey of PPSR projects, we found that most projects employ multiple mechanisms to ensure data quality and appropriate levels of validation. We created a framework of 18 mechanisms commonly employed by PPSR projects for ensuring data quality, based on direct experience of the authors and a review of the survey data, noting two categories of sources of error (protocols, participants) and three potential intervention points (before, during and after participation), which can be used to guide project design.**

***Keywords*-citizen science; data quality; data validation**

I. Introduction

Citizen science is a form of research collaboration in which professional scientists engage with members of the public in order to accomplish scientific research [1]. Due to the wide variability in skills and expertise between contributors, issues of data quality often rise to the forefront in considering the validity of this research. Data quality refers to the fitness of data for an intended purpose, and establishing data quality typically involves a multifaceted evaluation of states such as completeness, validity, consistency, precision, and accuracy. Although numerous examples of excellent results have been published (e.g., [4], [5]), the ability to generate scientifically rigorous results depends on the design of the research and participation tasks [3]. In this paper, we discuss the results of a survey reporting upon the mechanisms used by citizen science projects to ensure scientific quality. We then review and categorize additional mechanisms, inductively expanding from the survey and direct experience with citizen science projects to include additional options that logically follow from observed mechanisms. Finally, we present a framework for data validation and quality improvement.

II. Background

Research across disciplines conducted following the PPSR model typically focuses on either data collection, such as eBird and Monarch Watch, or data processing, such as Stardust@Home and Galaxy Zoo. Monitoring and observation oriented projects are centered on collecting data from contributors, often at larger temporal and geographic scales than are otherwise possible, while data processing projects leverage human perceptual capacities and problem-solving skills to accomplish analysis tasks that are currently too difficult for computers, spanning a wide variety of task types [2]. The design of research within each of these types is highly variable, but have common challenges in ensuring quality data.

These projects are increasingly enabled by and take advantage of information and communication technologies to advance scientific research [9]. They are often considered a type of crowdsourcing, a term referring to a set of distributed production models that make an open call for contributions from a large, undefined network of people [6]. Like other forms of crowdsourcing, most citizen science relies on adequately large numbers of contributors [7], along with a variety of mechanisms to ensure valid research results. Although an increasing variety of projects are using similar methods for improving data quality and research validity, no review of these mechanisms is readily available for reference. Our thesis is that while most scientists' initial impulse may be to employ the data quality methods standard in their field, when designing citizen science projects, a different approach to ensuring data quality may be necessary, taking into consideration the scale of participation and expectations around contributors' skills.

III. Methods

We used two sequential methods, starting with a survey of citizen science projects, and then inductively developing a framework based on the survey results as well as direct observation and participation by the authors.

A. Survey Instrument

A survey instrument was composed to directly elicit selected descriptive characteristics of projects. It was presented as a two-part questionnaire: first, a brief project profile and second, a separate, lengthier survey.

The first portion of the questionnaire was a project profile, allowing projects to opt-in for listing on several cooperating websites that provide listings of citizen science projects, and update existing project profiles based on data provided with the sampling frame or create a new project profile. The project profile included 23 items that would be considered useful by potential participants; the second portion of the questionnaire was the project survey, which asked for additional details in several categories. The full survey included 57 items, with free-response spaces for each structured item. There were no required fields, so each item had a variable response rate. The items covered several categories, but those reported in this paper focused on data validation methods.

B. Sample

The sampling frame was composed of projects listed on Cornell Lab of Ornithology's citizen science email list and in the now-defunct Canadian Citizen Science Network. These are the most comprehensive sources of contacts for North American citizen science projects. Approximately 60 additional contacts were manually mined from the online community directory at http://www.scienceforcitizens.net to extend the disciplinary diversity of the sample.

These sources provided a combined set of approximately 840 contacts after removing duplicates and bad addresses. These contacts are individuals who had self-identified as responsible for or interested in the management of citizen science projects. Approximately 280 projects were identified in this process, and another 560 individuals who may be connected with additional projects were also invited to participate.

C. Response Rate

In response to approximately 840 emailed requests for participation, 128 project profiles were created or updated. 73 surveys were initiated and 63 fully completed, for a participation rate of 15% and a response rate of approximately 8%. The surveys and profiles were combined for analysis.

The response rate is low, though not atypical for such a survey. However, it should be noted that contacts were asked to report on projects, and the number of projects is smaller than the number of contacts, meaning that the response rate for projects (our unit of analysis) is better than it appears. As noted above, we were able to identify approximately 280 projects, which would lead to a response rate of about 22% rather than 8%; the actual response rate lies somewhere in between these two figures.

Most of the responses came from small-to-medium sized projects, based in the United States, with several Canadian projects reporting along with two from the UK. The sample therefore best represents medium-sized North American citizen science projects, and nearly all responding projects are of the monitoring and observation types. Additionally, there is likely selection bias due to the associated project directory listing option, which means that projects not interested in active or general recruitment might be less likely to seek publicity. However, despite these limitations, we believe that the resulting sample is generally representative of the population of citizen science projects. Independent review of the response pool characteristics by staff at the Cornell Lab of Ornithology, who have conducted numerous similar surveys, suggested that the responses provide a fairly representative sample of the larger community.

D. Additional Review

Following the survey, the authors met to aggregate our notes from independent observation of citizen science projects for which we have direct knowledge as a result of involvement as researchers (over 50 projects). This allowed us to inductively elaborate on the mechanisms specified in the survey to generate a more complete list of mechanisms that we have observed in use, and based on these, to identify additional potentially useful mechanisms. We then considered the characteristics of these mechanisms to generate a framework for guiding selection of validation methods and data quality improvement.

IV. Results

We briefly describe our survey sample and responses to the items on the subject of data validation, and then discuss additional mechanisms.

A. Survey

1) Resources in Citizen Science Project Sample: The responding projects reported between zero and over 50 paid full-time equivalent employees (FTEs). Of 50 respondents for this item, the average number of FTEs was 2.4 but the median was one. Several noted that this allocation of staffing was spread across numerous individuals, each contributing only a small fraction of their time. Annual budgets ranged from $125 to $1,000,000 (USD or equivalent); 43 projects responded with estimated annual budgets, with an average of $105,000 but with a median of $35,000 and a mode of $20,000, indicating substantial variability in available resources. Projects identified up to five different funding sources employed to meet their expenses, most commonly drawing upon grants and in-kind contributions, typically of staff time, as well as private donations.

52 projects included the year founded in their responses. Responding projects were widely variable with respect to the age or duration of the project. A few projects were not yet

Table I
VALIDATION METHODS REPORTED

Method	n	Percentage
Expert review	46	77%
Photo submissions	24	40%
Paper data sheets submitted along with online entry	20	33%
Replication or rating, by multiple participants	14	23%
QA/QC training program	13	22%
Automatic filtering of unusual reports	11	18%
Uniform equipment	9	15%
Validation planned but not yet implemented	5	8%
Replication or rating, by the same participant	2	3%
Rating of established control items	2	3%
None	2	3%
Not sure/don't know	2	3%

Table II
COMBINATIONS OF MECHANISMS REPORTED

Methods	n	Percentage
Single method	10	17%
Multiple methods, up to 5 (average of 2.5)	45	75%
Expert review + Automatic filtering	11	18%
Expert review + Paper data sheets	10	17%
Expert review + Photos	14	23%
Expert review + Photos + Paper data sheets	6	10%
Expert review + Replication, multiple	10	17%

operational, and one was 100 years old. The average age of currently operational projects is 13 years, while the median is 9 years and the mode is 2 years.

2) Data Validation Responses: There were 60 responses to the question, "What methods of validation or quality control are used? Please check all that apply," (Table I.) The most common mechanism (of those provided in the list) for data validation is expert review, followed by photo submissions. The submission of photos as a mode of validation for data contributions is an interesting choice, given the challenges of processing, storing, and archiving photos. Without additional infrastructure (e.g., permitting online image identifications or classifications) this mechanism is not likely to be well suited for large-scale projects.

A surprising number of projects (33%) also require the submission of paper data sheets along with online data submissions, which may seem counter-intuitive if the assumption is that online data collection obviates paper-based data collection. In interviews for a related research project, however, project organizers indicated that their online databases do not accommodate the full range of data or details that are collected in the field, or that the paper data sheets are randomly sampled and verified against the online records to ensure accuracy of data entry.

The majority of projects employ multiple mechanisms to ensure data quality, and the most common combinations include expert review along with additional documentation of observations (Table II). This reflects in part the dominance of data collection as the primary task for contributors, but also concerns over accurate identification, for example, of species or phenophases (life cycles of plants and animals.)

Table III
RELATIONSHIP BETWEEN RESOURCES AND VALIDATION METHODS

Method	Variable	r^2
Uniform equipment	Staff	0.15
Uniform equipment	Budget	0.19
QA/QC program	Budget	0.14
Photo submissions	Budget	-0.10
Expert review	Budget	-0.29
Paper data sheets	Budget	-0.21
Number of methods	Staff	0.11
Number of methods	Budget	-0.15

3) Project Resources and Data Validation Choices: Using FTEs and annual budget as measures of organizational or institutional commitment to a project, we would expect that resource-rich projects might invest more in data quality and validation mechanisms. We examine correlations between these variables to examine this general prediction in several ways. For example, we might expect the use of uniform or calibrated equipment to be positively related to both staff and budget, as it would logically require more funding and people to manage appropriately, which is empirically supported (see Table III.) QA/QC programs are also positively correlated with budget, as they are typically costly to implement at any scale. On the other hand, use of paper data sheets and experts for data review are negatively correlated with funding.

Perhaps most surprising, the number of validation methods and budget are negatively correlated, while the number of validation methods and staff are positively correlated. This violates the assumption that projects with more financial resources might apply more mechanisms to ensure data quality, although more staffing would suggest more supervisory capacity. If we assume that projects with larger budgets also have larger numbers of contributors, it seems clear that scalability may be constrained with respect to the number of validation mechanisms, in addition to the degree of human involvement required. It is also possible that more funding for larger scale projects leads to fewer but more sophisticated mechanisms. This observation has interesting implications for the design of projects for large-scale participation.

4) Other Validation Methods: We also received a few responses to the open-ended question, "Please describe any additional validation methods used in your project." These responses are listed below.

- Instruments calibrated annually at a Major Observatory
- Evaluating observer reliability with evaluation program
- Participants known and qualified
- Participants send in samples for identification
- Know credibility of contributors using password access
- Measurements
- Done by staff of the project

- Test at all training classes
- Scientific QA/QC at all levels; peer-review
- Results are reviewed by stakeholder committee and project teams
- Manual filtering of unusual reports
- Piloting efforts to have expert ecologists ground-truth participant submissions
- On-line data entry subject to peer and expert review and analysis
- Some are built into data entry system (set ranges for location, temperature, water quality measures, completeness, etc.)

Three observations can be gleaned from this list. First, several projects depend on personal knowledge of contributing individuals in order to feel comfortable with data quality. This is not a particularly strong method from a scientific standpoint, but is understandably relevant with respect to practice. Second, most of the comments refer to the form of expert review that is employed, further reinforcing the perceived value of expertise in ensuring data quality. Notably, professional ecologists' re-use of data may be affected not only by trust in the data collectors' ability to collect quality data, but also by the comprehensibility of how the data were collected (e.g., the protocols used and the local context of data collection) as described in metadata [10]. Third, the final comment in the list is relevant to most projects with online data entry, though often overlooked: the reporting interface and associated error checking provides an initial and important form of data verification.

B. Review of Additional Mechanisms

In discussions, we inductively generated a more comprehensive list of data quality and research validation mechanisms (Table IV) that can be usefully categorized according to the point in the research process at which quality is being addressed and the presumed source of error. In addition to error, data may have accuracy and precision characteristics along several axes, such as taxonomic, spatial, temporal, and other attributes. One can therefore think about data validation mechanisms with respect to the error that is being prevented: malfeasance (gaming the system), inexperience, data entry errors, etc.

These observations suggest several additional questions to consider during the design or review of data quality and validation mechanisms.

- Is the source of error presumed to originate with the contributors or the protocol?
- How do mechanisms address the presumed sources of error?
- When are mechanisms active?
- Is data quality ensured before participants' contributions, during, or afterward?
- How transparent are data review processes and outcomes?
- Who sees the outcomes of data review?
- How much judgment is required to make a decision about accuracy, and how much data dimensionality is evaluated?
- How much of the data will be reviewed?

As noted in the prior section, the design of data collection and online submission forms is an important mechanism for *a priori* data quality management. We do not elaborate on this in our list of mechanisms, but suggest a few basic characteristics to consider in the design of data entry forms. Online forms can employ any of the following to improve data submission quality: controlled vocabularies, machine timestamps, required fields, field formatting, and set ranges for data entry. Professional assistance with data entry interface design may also prove valuable, and in the context of citizen science, considering the effects of incentive systems on data submissions is also important.

C. Framework for Citizen Science Data Quality

We now introduce the framework resulting from our data collection and analysis. This framework documents a suite of data quality and validation mechanisms, showing the point in the research process when the method is applied (before, during, or after data collection), and briefly describes relevant variations and details related to each mechanism in Table IV.

A general observation about the application of data quality mechanisms in citizen science relates to the form of contributions that are solicited. These contributions typically take the form of either data collection or data processing. A wide variety of mechanisms are used in projects that focus on collecting data; our survey results primarily reflect this form of participation. In projects that rely on contributions to data processing, the vast majority require classification, coding or annotating images or text by multiple individuals, using inter-rater reliability as an indicator of quality or the need for expert review.

V. Discussion

A. Survey

When we consider the top mechanisms reported by citizen science projects, we find that most of the reported methods fall into the framework categories of contributor and contribution process; few methods that were reported were in fact focused on the data outcomes. We note, of course, that our survey items did not specifically elicit mechanisms that address the data validation or quality after data collection.

Also notable, however, is that out of the free text responses, only one referred to data, focusing on the data entry end. Further attention to the way that citizen science data are created–not just the protocols for participation, but qualities such as format and precision, for example when working with geographic data–represents an area of need for this domain of practice. Missing from this picture are the

Table IV
FRAMEWORK OF OPTIONS FOR DATA QUALITY.

Mechanism	**Process**	**Source of error**	**Types and details**
Quality assurance project plans	Before	Protocols	Standard operating procedure in some disciplines
Repeated samples/tasks	During	Protocols	*By multiple participants:* common crowdsourcing approach, e.g. duplication of input *By the same participant:* usually site-based, over time; participant error may be replicated or corrected *By experts:* single site calibration by experts for multi-site data collection
Participant tasks involving control items	During	Protocols	Contributed data are compared to known states for both image recognition tasks and monitoring multiple permanent plots
Uniform or calibrated equipment	During	Protocols	Used when measurements are taken; cost/scale tradeoff; who bears cost?
Personal knowledge of participant skills/expertise	All	Participants	Does not scale well; hard to demonstrate reliability; surveys may be useful
Participant training	Before, During	Participants	*Initial:* cost depends on scale, mode of delivery; barrier to participation *Ongoing:* high cost, most practical for localized projects *Formal QA/QC:* high cost, most often in water quality projects
Participant testing	Before, During	Participants	*Following training:* often prerequisite to data acceptance *Pre/test-retest procedures:* may impact participant retention
Rating participant performance	During, After	Participants	*Unknown to participant:* may require more data contributed by or additional info about participant
Filtering of unusual reports	During, After	Participants	*Known to participant:* care required; can de/motivate performance *Automatically:* algorithmic identification of outliers *Manually:* sorting and filtering by researchers, often with spreadsheets
Contacting participants about unusual reports	After	Participants	Potential to alienate/educate contributors
Automatic recognition techniques	After	Participants	Computer science techniques for image/text processing, e.g. for tagging species data for verification
Expert review	After	Participants	*By professionals:* usually scientists associated with the project *By experienced contributors:* long-term volunteers or recruited experts *By multiple parties:* any combination of reviewers, including peer review
Paper data sheets submitted in addition to online entry	During	Protocols, Participants	Useful for extended details not accommodated in databases and verifying accurate data entry
Digital vouchers	During	Protocols, Participants	*Photos:* with or without EXIF data, to make or verify species identifications *Audio:* some sounds not recordable with smartphones, e.g., cricket calls *Museum/herbarium specimens/archives*
Data triangulation	After	Protocols, Participants	Corroboration from other data sources, e.g., remote sensing data, qualitative data, historical trend data
Data normalization	After	Protocols, Participants	Standard and advanced statistical techniques
Data mining	After	Protocols	Computer science techniques, requires very large data sets
Data quality documentation	After	Protocols, Participants	Provide metadata about what mechanism(s) were used

methods that can be employed to work with citizen science data in the analysis process to ensure validity. We currently find relatively few examples of good methods to deal with the data that are collected on a large scale but are spatially or temporally incomplete, geographically biased, or involve some level of known error.

One solution is applying data mining methods from computer science, or collaboration with researchers in this area. There are some excellent examples of analytic approaches to ensuring validity emerging (e.g. [8]), but additional work in this area is clearly needed. The primary point of departure from prior research methods with which researchers are comfortable are that these data are differently generated, and therefore need different analytic considerations.

Another point that the survey raised is the scalability of validation methods. The correlations between project characteristics and chosen validation mechanisms seem to clearly relate to the human and fiscal resources available for each project. Most methods, including automatic data filtering or recognition, observation or tasks involving control items, or replication of tasks by multiple contributors, were not strongly correlated with either staffing or budget, suggesting that they are instead determined by the research at hand. Increased budget and staffing yield projects with fewer instances of labor-intensive validation methods, less reliance on paper data sheets as a backup to electronic records, and fewer methods overall.

If we assume that projects receiving more funding also

involve more contributors, this suggests that methods such as expert review and collection of paper forms do not scale well, leading projects with adequate resources to employ other methods of data validation. This has substantive implications for the design of projects based on their expected contribution base and growth trajectory. It suggests that projects may need to plan different quality management techniques based on the projected growth and resulting size of the data set.

B. Framework

While we can make no claims as to the exhaustiveness of our list of mechanisms for data validation and quality in citizen science, we believe it presents a more complete selection of options than has previously been assembled. Depending on the implementation, each of these methods may address sources of error from the participants, protocols, and data. The rationale for the choices of mechanisms should be determined by the nature of the research at hand, but our recommendation is based on the observation that most projects address only error introduced by the contributors, and secondarily, error resulting from the protocol itself.

We believe that additional data quality issues can be addressed by thinking more holistically about the nature of error in citizen science, and the stages of research in which error can be corrected or ameliorated. Focusing on the use of carefully designed data entry forms can improve the quality of data, and data mining techniques for analysis of large-scale but biased or incomplete data can be valuable for improving the validity of resulting interpretations.

C. Future Work

Much remains to be done in the area of data validation and quality in citizen science. As our results and discussion reveal, most projects show greater concern over the lack of contributor expertise than the lack of analysis methods suited to the type of data generated in citizen science. It is clear that there is much work to be done with respect to analysis methods that are appropriate for these data; this is a major challenge, as many projects lack resources to engage computer science researchers and statisticians to help develop suitable data mining algorithms and models.

In addition, evaluation of the efficacy of data validation and quality mechanisms is a logical next step. Evaluation of the usefulness of our framework for the design of citizen science projects would demonstrate its value, which could be achieved by further developing it into a QA/QC planning and evaluation tool. Finally, regardless of selected data quality assurance mechanisms, citizen science projects must ensure that they adequately document these choices and convey the mechanisms employed along with the data they disseminate. This is necessary not only for adequately satisfying the expectations of peer review processes for publication, but also for data re-use.

VI. Conclusion

This paper discussed some issues around data validation and improving data quality in citizen science projects. Most projects employ multiple mechanisms to ensure quality, with these selection driven in part by available resources and scale of operations, which suggests that the scalability of data validation mechanisms is an important consideration for citizen science project planning and development. While there is substantial work remaining to advance the state of the art in citizen science data validation and quality improvement, this paper contributes new insights into existing practices and a framework to guide project design choices.

Acknowledgment

The authors thank the members of the DataONE working group on Public Participation in Scientific Research for discussions contributing to this paper. This work is partially supported by US National Science Foundation Grants 09-43049 and 11-11107.

References

[1] Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K., and Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59(11), 977–984.

[2] Cho, A., and Clery, D. (2009). Astronomy Hits the Big Time. *Science*, 323(5912), 332.

[3] Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience*, 58(3), 192–107.

[4] Delaney, D.G., Sperling, C.D., Adams, C.S. and Leung, B. (2008). Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions*, 10, 117–128.

[5] Galloway, A.W.E., Tudor, M.T., and Vander Haegen, W.M. (2006). The reliability of citizen science: A case study of Oregon white oak stand surveys. *Wildlife Society Bulletin* 34:1425–1429.

[6] Howe, J. (2008). *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Business.

[7] Kelling, S., W.M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. 2009. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, 59:613–620.

[8] Munson, M.A., Caruana, R., Fink, D., Hochachka, W.M., Iliff, M., Rosenberg, K.V., Sheldon, D., Sullivan, B.L., Wood, C., and Kelling, S. (2010). A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution*, 1(3), 263–273.

[9] Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution*, 24, 467–471.

[10] Zimmerman, A.S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science Technology & Human Values*, 33:631–652.