Managing Data Quality in Observational Citizen Science

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

S. Andrew Sheppard

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Loren Terveen

December, 2017

Acknowledgements

This thesis would not have been possible without the input and support of several of my colleagues and mentors over the years.

I thank my advisor Loren Terveen, who bore with me while I found my research footing and adapted to my non-traditional approach. I am grateful to David DeMuth and Linda Kingery, who introduced me to citizen science and supported my entry into graduate research. Thank you to Brian Fischer, who provided the much needed flexibility in my work schedule to be able to complete this thesis.

The GroupLens research lab was an exciting and supportive environment from which to launch my research projects. Thank you to John Riedl, Shilad Sen, and Reid Priedhorsky for welcoming me into the lab and mentoring me in my early years. Thanks to Aaron Halfaker, Fernando Torre, Mikhil Masli, and Jacob Thebault-Spieker for your friendships and thoughtful discussions around the lunch table.

I thank Wayne Goeken, Danni Halvorson, and other River Watch affiliates for their ideas and patience as I studied their program. I also thank Julian Turner and Noah Newman for their insight into the inner workings of CoCoRaHS. And a special thank you to Andrea Wiggins, who helped me connect with the larger citizen science community and focus my research contribution.

Thank you to past and present members of my committee, including Haiyi Zhu, Yuqing Ren, Daniel Keefe, and Brent Hecht. Their support and thoughtful critique helped to make this thesis more robust and engaging.

I thank my parents, Steve and Dawn, for instilling in me a lifelong curiosity and a determination to finish what I started. Finally, I thank my wife EunKyoung and daughter Vivian, for bearing through the paper deadlines and unexpected delays. I know you are even more excited than I am that the end is finally here.

Abstract

Observational citizen science is an effective way to supplement the environmental datasets compiled by professional scientists. Involving volunteers in data collection has the added educational benefits of increased scientific awareness and local ownership of environmental concerns. This thesis provides an in-depth exploration of observational citizen science and the associated challenges and opportunities for HCI research. We focus on *data quality* as a key lens for understanding observational citizen science, and how it differs from the related domains of crowdsourcing, open collaboration, and volunteered geographic information.

In order to *understand* data quality, we performed a qualitative analysis of data quality assurance practices in *River Watch*, a regional water quality monitoring program. We found that data quality in River Watch is primarily maintained through universal adherence to standard operating procedures, rather than through a computable notion of "accuracy". We also found that rigorous data quality assurance practices appear to enhance rather than hinder the educational goals of the program participants.

In order to *measure* data quality, we conducted a quantitative analysis of *CoCoRaHS*, a multinational citizen science project for observing precipitation. Given the importance of long-term participation to data consumers, we focused on *volunteer retention* as our primary metric for data quality. Through survival analysis, we found that participant age is a significant predictor of retention. Compared to all other age groups, participants aged 60-70 are much more likely to sign up for CoCoRaHS, *and* to remain active for several years. We propose that the nature of the task can profoundly influence the types of participants attracted to a project.

In order to *improve* data quality, we derived a general workflow model for observational citizen science, drawing on our findings in River Watch, CoCoRaHS, and similar programs. We propose a data model for preserving provenance metadata that allows for ongoing data exchange between disparate technical systems and participant skill levels. We conclude with general principles that should be taken into consideration when designing systems and protocols for managing citizen science data.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Domain: Observational Citizen Science

Citizen science, a form of scientific collaboration that engages non-professionals in research, is an increasingly valuable means for collecting robust scientific data sets[1] . The benefits of including volunteers in scientific research include local ownership of environmental concerns, an increased "sense of place", and improved understanding of scientific matters, as well as enabling new research that was previously impossible. While amateur science is arguably the oldest form of science, its resurgence in the last two decades – fueled in a large part by the Internet – has profoundly transformed the scalability of the scientific enterprise.

Indeed, recent high-profile projects like Galaxy Zoo [82, 84] are largely virtual, taking advantage of the the web as infrastructure to facilitate data processing via human computation. These *virtual citizen science (VCS)* [119] projects share features with the crowdsourcing style of distributed work [39], in which large contributor bases are usually required for success. Given the large scale and technical focus, these projects are also attractive platforms for robust quantitative research in computer science and related domains.

However, the majority of citizen science projects still focus on collecting *observational data* of such phenomena as weather and precipitation, air and water quality, and

[1] This section was adapted from the introduction to [103].

species abundance and distribution. This form of citizen science, which we will hereafter refer to as *observational citizen science (OCS)*, has actively contributed to science and decision-making for over 100 years [26]. While most such projects now also take advantage of the web, field-based data collection tasks will always involve a physical component, leading to interesting tensions in the use of Information and Communication Technologies (ICT) for citizen science.

In contrast to the near universal scale of VCS, the geographic and human scale of OCS projects can vary widely. On one end of the spectrum are large-scale projects like the *Community Collaborative Rain, Hail & Snow Network (CoCoRaHS)* [14], a network of precipitation observers across North America; and eBird, which collects global bird abundance and distribution data [108]. Approximately 150,000 active participants generate around 5 million observations of birds per month for eBird, with a total data set size of around 140 million records in mid-2013. The data are used for scientific research as well as policy and land management decisions. CoCoRaHS data are requested nearly 3,000 times a day; as of early 2017, the largest of several data sets includes 35 million data points, each representing daily precipitation for a specific location. Notable data users for CoCoRaHS include the U.S. National Weather Service, National Climatic Data Center, and other climate and weather organizations.

On the other end are local scale projects, sometimes referred to as community-based monitoring [123]. These projects are often organized by local community leaders who are generally experts in the topic area, but may or may not have prior professional experience. Projects like *River Watch* [101] and Mountain Watch [118] usually have specific conservation or social justice goals. These projects benefit from working with support from, and exchanging data with, government entities and other related organizations. For example, River Watch data is shared with the Minnesota Pollution Control Agency and becomes a part of the official state record for the rivers monitored by the program. Mountain Watch data were used in testimony to lawmakers to demonstrate the value of maintaining Clean Air Act protections.

In addition to differences in scale, there are profound differences in task structure in OCS versus VCS. While VCS projects can often verify results by having multiple volunteers complete the same classification task, OCS data are much more individual,

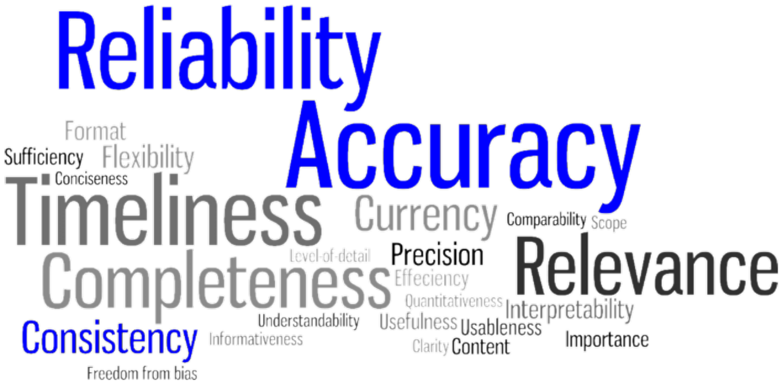


Figure 1.1: Commonly referenced "Dimensions of Data Quality"[112]

time and location-sensitive[2] . OCS data are often unique observations of changing natural phenomena; as such, they may not be directly verifiable. An eBird project leader described the act of observation as "the intersection of a person, a bird, a time, and a place," emphasizing that for these data, the observation is never repeatable.

How do these differences play out in practice? Observational citizen science involves a number of physical constraints that directly affect the contribution workflow, the structure of the data, and the metrics available to measure success. There is relatively little quantitative research evaluating these unique aspects of observational citizen science. This gap can potentially lead to faulty assumptions about the applicability of system design recommendations made in related domains [118].

This thesis provides a significant step toward closing that gap.

1.2 The Lens: Data Quality

There are a number of possible ways to characterize OCS, as well as to situate it among other application domains that receive more attention in the computer science literature. However, we propose that *data quality* provides a uniquely powerful lens with which to understand this domain, as well as the opportunities for computers to support it.

Data quality is notoriously domain-dependent, and it is difficult to define a universal, easily computed operationalization. Figure 1.1 shows a number of common dimensions

[2] The location-dependent nature of field observations can also lead to complex privacy issues, which fall outside of the scope of this thesis.

of quality, derived from a literature review by Wand and Wang [112]. While "accuracy" is the most commonly referenced dimension, it is hard to operationalize without understanding the context in which it is being applied. Most research disciplines thus define data quality simply as "fitness for its intended purpose" [43].

In OCS, data quality is often critical to achieving project goals, especially when the data may be used by resource managers and government agencies to make assessments of the environment and resulting policy decisions [103, 121, 66]. Without an assurance of data quality, domain experts may question the quality of the information, since it is maintained by volunteers with unknown and varying levels of knowledge and skill. Further, OCS contributions consist of highly individualized data that is difficult to independently verify. Data quality in OCS is often maintained via training and standard protocols ("get it right the first time"), rather than the iterative improvement workflows in Wikis and similar systems that deal in externally verifiable facts.

"Improving data quality" is the goal of the technical approach proposed in the last chapters of this thesis. But in order to improve quality we first need to measure it, and in order to measure it we need to understand how it is operationalized. These questions touch at nearly every aspect of how OCS projects are organized. In fact, we argue that most of what is needed to fully understand an OCS project can be learned from knowledge of how that project manages data quality concerns.

Thus, data quality is the primary lens that connects all of the chapters in this thesis.

1.3 Related Domains

This thesis builds on related work in the domains of *crowdsourcing*, *open collaboration* (*peer production*), and *volunteered geographic information.* At a high level, all of these domains involve community-maintained artifacts of lasting value created via shared web-based repositories [6, 17]. For this reason, all of these domains regularly grapple with the same key questions addressed in this thesis: *How can we understand, measure, and improve the quality of data contributed by volunteers?* Thus, while there are important differences, these domains are useful for grounding and situating this work.

In some sense, *citizen science* overlaps with and could be considered just another

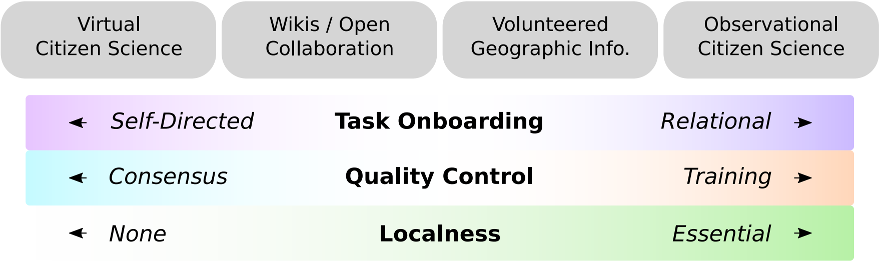


Figure 1.2: Spectrums of task design in citizen science and related domains.

term for these domains. However, relatively unique to citizen science is a focus on scientific inquiry (often toward environmental questions), and usually a more hierarchical approach than peer production models. Citizen science is most often used to describe projects with highly domain-specific tasks, defined by expert scientists and executed by volunteers. This unique relationship between professional scientists and volunteers in citizen science means that domain-specific research is needed to understand, for example, how participant motivations change over time [93].

Using data quality as a lens, we can plot these domains out along a number of spectrums that characterize the nature of project work in each domain. These spectrums are shown in Figure 1.2. The key differences relate to the structure of the data collection task, the primary quality control mechanism, and the "localness" of the data being collected [36].

1.3.1 Crowdsourcing (Virtual Citizen Science)

Crowdsourcing is a vague term, used to describe both paid microtasking work (e.g. Mechanical Turk [9, 91]), as well as large-scale unpaid image classification [60]. This second definition is often used interchangeably with virtual citizen science [119]. Throughout this thesis, we use crowdsourcing to refer primarily to large-scale VCS projects, and not paid microtask work.

VCS crowdsourcing projects typically do not require any local knowledge or in-depth training. The data collection task is typically defined in such a way that it can be performed in a short time by anyone with an internet browser. The archetypal VCS project is Galaxy Zoo, which was originally created to facilitate the classification of a large collection of galaxy images. Galaxy Zoo has since expanded into a number of similar classification projects, collectively known as the Zooniverse [87]. In Galaxy Zoo and most VCS projects, the key QC mechanism is consensus - when multiple people provide the same classification for the same image, it is likely to be correct.

1.3.2 Open Collaboration / Peer Production

Open collaboration, or peer production, is a process for involving a large number of contributors in the generation of useful content. One of the most prominent and successful examples of an open collaboration website is Wikipedia. A guiding principle of Wikipedia is that anyone in the world can edit nearly any article without establishing relationships with existing contributors [3] . While some articles are location-based, most involve general knowledge.

Like VCS, peer production relies on a form of consensus as the primary quality control mechanism. However, while consensus in VCS is well-defined (typically by project administrators as a numeric threshold), in Wikipedia it is more nuanced and relies on community norms and discussion.

Previous work on Wikipedia has confirmed that, like other forms of peer production, the majority of content is created by a minority of contributors. This has been characterized by the power law comparing the number of distinct authors per article [110] and the number of revisions per article [1]. While Kittur et al. noted a decrease over time in the relative proportion of contributions by power users [48], Priedhorsky et al. found that the contributions by top contributors are more likely to stick around long enough to be *viewed* and that this effect has been increasing over time [79].

1.3.3 Volunteered Geographic Information

Volunteered geographic information, or VGI [31], is the term most commonly used in the geospatial community to describe geo-referenced contributions from volunteers. While VGI systems come in many forms, the implementations most relevant to this discussion are *geowikis.* Like text wikis, geowikis rely on a collaborative refinement model. The most prominent geowiki is OpenStreetMap [32].

[3] In practice, the Wikipedia community has evolved into a hierarchy of administrators and core contributors with more power than other users [10].

OpenStreetMap is a community-based effort, and there are numerous real-world opportunities for participants to meet each other. Hristova et al. found that mapping parties have a positive impact on both recruitment and retention [41]. Palen et al. found that key crisis events served as a catalyst for better community organization [74]. Hristova et al. found that the unequal contributions found in the power law are somewhat attenuated within local communities of OpenStreetMap participants working together [40].

1.3.4 Observational Citizen Science

Wiggins and Crowston list five exhaustive categories of citizen science projects [119]. For the purposes of this discussion, we contrast *Virtual* projects with the four non-virtual categories (*Action*, *Conservation*, *Investigation*, and *Education*) which we consider variations of *observational citizen science.* Depending on the context, OCS can also be referred to as participatory sensing, participatory action research, volunteer monitoring, or community based monitoring [16]. We use *observational* to draw attention to two key aspects: the structure of the task - which requires physical presence at the phenomena being observed, and the structure of the data - which is typically a time series of observations.

Like VCS, observational citizen science projects are typically commissioned by scientific experts and executed by volunteers. These projects are often designed to engage non-professionals in a rich shared appreciation of the natural environment and the scientific method. This leads to research focusing on participant motivations and learning outcomes [5, 52, 61, 94, 125].

Other projects are more similar to open collaboration communities, in that citizen volunteers are treated not only as collectors of data but as fellow scientists [51] and participate in defining scientific goals as well as executing them [58]. Conrad and Hilchey [16] differentiate between a number of approaches to citizen science with varying levels of participation, ranging from traditional, centralized “consultative” programs to grass-roots, bottom-up “transformative” communities, while noting that the various approaches to citizen science are not necessarily mutually exclusive (c.f. [54]).

Regardless of the approach, OCS projects are centered around the generation of knowledge that is highly localized. This means that consensus-based QC mechanisms are often secondary to training, which often is structured around personal relationships with organizers and other volunteers. These characteristics provide a unique opportunity for CSCW and HCI to explore the design space that addresses the tensions, tradeoffs and synergy between the relational and data collection goals in OCS [7, 23, 42, 99, 53, 101].

1.4 Thesis Overview

The primary outcome of this research is a thorough examination of the unique characteristics of observational citizen science, with a goal of better informing the design of data management platforms for this domain. This includes three contributions:

- **Understanding data quality** via a qualitative description of data management practices and perceptions of quality control, derived via interviews and participant observation within the River Watch community-based monitoring program (Chapter 2)
- **Measuring data quality** via a quantitative model for participant retention that integrates demographics, geographical constraints, initial participation habits - statistically derived from millions of observations submitted to a national-level citizen science program (Chapter 3)
- **Improving data quality** via the design and implementation of a data model for retaining provenance, derived from an exhaustive overview of the general OCS workflow (Chapter 4)

In the remainder this thesis, I discuss each of these three studies before concluding with a summary of design implications and directions for future work.

Chapter 2

Understanding Data Quality: Task Structure in River Watch

2.1 Introduction

As discussed in the last chapter, citizen science and open collaboration are both effective ways to generate valuable repositories of information by spreading the effort out across communities of volunteers. However, domain experts in both fields often express concern about the quality of the contributed data.

Open collaboration communities tackle these quality concerns in different ways. One approach is developing a set of best practices that participants are encouraged to follow, such as Wikipedia policies [10]. These practices promote quality by formalizing the work needed to maintain it. However, rigorous enforcement of policies can become a detriment to community relationships and potentially come into tension with other project ideals, such as openness and ease of contribution [90, 122]. For example, Halfaker et al. have demonstrated that the increasingly strict enforcement of quality control by power users has become a deterrent to would-be newcomers to the community [33]. Among other things, this means that the coverage of various topics in Wikipedia is biased towards the interests of its most active contributors.

In contrast with Wikipedia, the level of rigor required in an observational citizen science community can depend on its emphasis. More interpretive approaches may accept that data will have lower quality, emphasizing instead awareness, participation, and educational aspects [97]. However, many projects do have a primary goal of collecting high-quality environmental data, in which case these projects must define precise and straightforward data collection protocols to encourage quality while not discouraging participation or subverting educational goals [66].

In this chapter, we investigate data quality concerns in the context of River Watch, an OCS community with both data collection and educational emphases based primarily in northwest Minnesota. We attempt to answer the following general research questions in that context.

RQ1. How is quality *defined* in an OCS community? How is it measured and operationalized – both by volunteers within the community and by "experts" on the periphery?

RQ2. How is quality *maintained* – i.e. how is community work organized to maintain quality? What kinds of errors might occur and how are they addressed?

RQ3. How is quality *balanced* among other community goals? Can the strictness that high quality requires come into tension with those goals?

To answer these questions, we interviewed River Watch participants, ranging from high school students to the staff of local government agencies. In the rest of this chapter, we will first discuss related work and the background of River Watch, then present results from the interviews and implications for the measurement and improvement of data quality in OCS communities like River Watch.

2.2 Related Work

2.2.1 Defining Quality

With the abundance of data and information resources available to organizations, quality is of critical importance. Low quality data can lead to inefficiencies and poor decisions that can have severe financial and social impacts. Thus, the field of data and information quality[1] is an active area of research [112]. While data quality may seem to be an

[1] Throughout this thesis, "data quality" is used to refer to both data and information quality.

intuitive concept (e.g. "correctness"), it is impossible to propose a precise, universally meaningful definition [106, 112, 114]. The concept of data quality must instead be understood in the context of the domain in which it is being applied and the purpose for the data.

As discussed in Chapter 1, Wand and Wang identify a number of common dimensions used in data quality research [112]. These consist both of *internal* dimensions (such as accuracy, reliability, and consistency) and *external* dimensions (such as timeliness, relevance, and interpretability). Internal dimensions are focused on the design and processes for producing data products, and can be understood from the perspective of the individuals responsible for producing the data. External dimensions are focused on the uses (or usefulness) of data, and can be understood from the perspective of data consumers.

Strong et al. assert that quality must be understood primarily from the perspective of the consumer [105]. Wang and Strong investigated quality from this external perspective, by conducting surveys of consumers of data [113]. In this chapter, we focus primarily on the processes and individuals responsible for collecting data. Therefore, we discuss the contextual definition of data quality primarily along the line of the *internal* dimensions discussed by Wand and Wang. We justify this approach based on the particulars of this domain, as discussed below.

2.2.2 Quality in Wikipedia

In Wikipedia, quality has investigated from the external perspective, as in Giles' well-known study [29] that investigated the quality of articles by having independent reviewers evaluate them. However, because in open collaboration the line between producer and consumer is blurred [106], it is also valuable to study data quality from the perspective of the community itself.

Stvilia et al. [106] analyzed Wikipedia logs to understand how work is organized to promote quality, with the goal of finding how the Wikipedia community understands quality, and what quality assurance practices are present. We intend to answer similar questions in River Watch, and this work is in part a response to the call to extend their work to other open collaboration communities. A challenge in doing so is that the quality assurance practices in River Watch largely happen offline, so there is not currently a robust repository of data available to track those practices quantitatively as there is in Wikipedia. We propose such a system in Chapter 4.

Our approach is more similar to that of Riehle [90], who investigated quality (among other concerns) by interviewing key contributors to Wikipedia. A common theme in Riehle and in Stvilia et al. is the discussion of various identifiable roles for maintaining quality in Wikipedia. Similarly, we have identified a number of specific roles in River Watch and have made a point to interview representatives from each of those roles.

There have also been various attempts to compute quality in Wikipedia automatically, for example in Halfaker et al.[34]. The metric proposed in that work relies on the key assumption that the editors of Wikipedia are the best judges of the quality of contributions to Wikipedia. Similarly, our work assumes that quality in River Watch is best understood by those actively involved in the program, and the quality improvements we propose are designed to leverage their existing knowledge.

2.2.3 Quality in OCS

The quality of data in OCS programs has been evaluated from both external and internal perspectives. Nicholson et al. compared volunteer and professionally-collected monitoring data statistically, finding that the data quality is comparable for certain parameters [66]. Hochachka et al. describe the robust statistical techniques used to accurately estimate species distributions based on observations submitted by eBird volunteers [37]. Sullivan et. al describe how these filters in combination with robust human computation processes help ensure a high quality dataset that can be used by a number of external agencies [107]. By contrast, Nerbonne and Nelson [63] operationalized data quality in community-based monitoring groups based on an understanding of the procedures used by each group and the quality assurance plans in place. We found this latter, primarily internal approach to be more valuable in the context of River Watch.

Citizen science has long been recognized for its educational benefits [66, 97], and is often incorporated in school settings as part of the curriculum or as an extra-curricular activity. Depending on the approach, there can be a tension between emphasizing data quality and promoting educational goals. River Watch prioritizes both education and data quality, so in this study we explore how those priorities interact.

2.2.4 About River Watch

River Watch is a community-based monitoring program that collects water quality data in the basin of the Red River of the North[2] . The program started in 1995 with 4 high schools, and now consists of 35 schools monitoring more than 200 sites on the Red River and its tributaries. As stated on the website[3] , River Watch is designed to help students:

- Develop a "sense of place" and connection to the local watershed.
- Learn field-based physical and biological ambient water quality monitoring skills.
- Establish connections to scientists engaged in watershed science.
- Become active contributors to the scientific community.
- Develop workplace skills.
- Provide important services to their local watershed.

While these educational aspects of River Watch are important, the data collected by the schools are also a valuable resource for the Minnesota Pollution Control Agency (MPCA) and local watershed districts in the basin. Due to limited staff at these agencies, many of the sites monitored by River Watch would not otherwise be monitored. Data collected through River Watch are combined with data collected by professional monitors and used to determine whether a particular water body is impaired, and even to identify specific sources of pollution. Because of the potential consequences of this process, it is critical that the data be of good quality.

There are a wide variety of water quality[4] parameters measured by River Watch teams, and a detailed discussion is beyond the scope of this chapter. However, they can be broadly categorized in the following different types:

[2] "River Watch" is not a unique name for a program of this kind. There are similarly named programs around the world and even within the state of Minnesota [21]. However, for the purposes of this thesis, River Watch refers specifically to the citizen monitoring program operating in the Red River basin under the auspices of the International Water Institute.

[3] `http://iwinst.org/education`

[4] Note that *water* quality is an environmental concept altogether different than the issues of data quality being addressed in this chapter. In effect, we are studying the quality of water quality data.

- Subjective descriptions of the state of the river (appearance, recreation suitability, field observations)
- Measurements taken by lowering an electronic probe into the water (pH, Dissolved Oxygen)
- Measurements taken by sending a sample of the water to a lab for analysis. This process presents a number of additional quality concerns, including chain-of-custody forms, holding times, and lab quality control. River Watch data is primarily field-based, so we will not address lab concerns in this chapter.

2.3 Methodology

I have been working with the project for about ten years as the designer and maintainer of the project's central data management platform [5] . Therefore, a practical goal of this project was to identify quality concerns in the River Watch data collection process, in order to address them through technological means where possible. We first explored how data quality is operationalized by the River Watch participants and leadership, by conducting a qualitative study of the River Watch program, involving both observation and interviews, as in Ribes and Finholt [89].

River Watch does not consist only of individual teams collecting data in isolation and submitting it to a central repository. It is a geographically localized community project, and there are multiple occasions for interaction with other participants. I attended at least one of each of the following activities:

- Annual River Watch Forums, in which representatives from nearly all of the schools gather for a day to share ideas and present posters detailing the lessons they gained from the previous monitoring year.
- Monthly curriculum committee and student leadership meetings, involving a subset of River Watch schools and coordinators meeting to set educational goals.
- Annual training and certification for River Watch and other water quality monitoring programs in the Red River basin.

[5] `https://river.watch/`

- Monitoring Advisory Committee meetings, where coordinators meet with agency staff and other water quality monitors to discuss and refine common procedures.

- Data collection field trips with River Watch schools and coordinators.

I also interviewed 15 participants from a number of distinct roles (see Table 2.1). Most of the individuals interviewed are actively involved in River Watch or otherwise supportive of it. We discuss the results of our study below.

	N	Description
Students	7	Teams of high school students form the primary volunteers in River Watch, and are usually responsible for doing the actual monitoring activities.
Teachers	3	Each River Watch team has an adult adviser, most often a natural science teacher. Teachers are generally responsible for making sure the students “stay on task” during monitoring trips, and work to integrate River Watch into curriculum goals where possible.
Coordinators	3	River Watch coordinators form the primary leadership of the program, and are responsible for training volunteers and promoting the program. While they are quite knowledgeable in their own right, they work closely with agency staff to define monitoring goals and procedures. They also work closely with schools and local universities to define and promote the educational and scientific goals of the program. Coordinators play a key “gatekeeper” role in that they are directly accountable to the MPCA for ensuring the quality of the data collected.
Agency Staff	2	MPCA and watershed district staff are responsible for setting monitoring goals and formalizing standard operating procedures. These are the “experts” in the domain, and while those we interviewed are supportive of River Watch they are not involved in the day-to-day operation of it.

Table 2.1: Roles in River Watch

2.4 Results

2.4.1 How is quality defined?

RQ1. How is quality defined in an OCS community? How is it measured and operationalized – both by volunteers within the community and by "experts" on the periphery?

Before designing systems for improving data quality, it is critical to understand how quality is understood by those responsible for maintaining it [112]. Our initial goal in studying this issue was to define metrics that could be computed on individual pieces of data. However, through the interviews it became clear that a "data-centric" approach to quality was incomplete.

At the start of each interview, participants were asked to define "data quality" in their own words. While the question was intended with the assumption that "quality" is a property of "data", a plurality of participants responded by describing quality assurance practices in River Watch, essentially answering the question "how is data quality maintained?" (our second RQ) rather than "what is quality?". As one teacher put it:

> "When I think of data quality, I think of protocols – someone monitoring the students while they're conducting the tests so that those protocols are followed."

This participant went on to describe other aspects of data quality - data integrity, accurate transcription, and correct interpretation, before concluding:

> "Data quality seems to me then to encompass a wide range of things – from the sampling procedures all the way to explaining what it means; and doing it as accurately as you can."

Quality in River Watch is operationalized in the methods used to collect and understand the data, rather than merely being a computable attribute of the data. Thus, the measurement of quality in River Watch is inseparable from the actions used to maintain it. Quality is not a static attribute but a set of actions; it is not just an adjective but also a verb.

Dimensions of Data Quality. This perspective can be further examined by exploring data quality in the context of the dimensions identified by Wand et al. [112]. While these dimensions were not used explicitly to frame the interview questions, they serve as useful lenses for discussing how data quality is perceived. *Reliability*, *accuracy*, and *consistency* in particular were concepts that came up often. The definitions of some of these terms are not universal (even within data quality research), so it is important to define what these mean in the context of River Watch.

Reliability

Reliability refers to the trustworthiness of the data, and its usefulness for answering pertinent questions about the watershed[6] . Implicit in the term is that there is someone that is relying on the data. River Watch data is relied on by the MPCA and others to make assessments of water quality and to make decisions that can cost millions of dollars. One teacher in particular described how their team had identified severe issues with turbidity in their watershed. The reliability of their data was particularly important due to the potential political implications of their findings.

> "If you're trying to tell people that we have a lot of sediment, and a lot of runoff – and maybe we don't have as much as what we're saying – that might create some hard feelings. Because a lot of it gets pointed at the agriculture and if our data would not be correct, maybe we're pointing fingers where we shouldn't be. But we feel that our data is consistent, and has been correct in that particular area."

Unreliable data would hurt the reputation of the program, as another teacher noted.

> "If it's going to be reliable, and River Watch is to be respected among the professionals, I think data quality has to be right at the top. Because you want credibility with other professionals."

On a more practical level, the longevity of the program is at stake, according to one coordinator:

[6] Strictly speaking, reliability is typically defined in data quality research as an internal dimension describing the process for generating data. In this sense, it is closely related to consistency, as discussed later in this chapter.

> "If you don't have good quality data why bother even doing the project – that was the selling point for the funders from watershed districts to support the program."

Participants were also generally concerned with the effect poor data could have on the conclusions drawn by others. As one student put it:

> "Other people look at it, and base their research and stuff off our data we collect, so it kind of has to be right, or else their stuff will be messed up."

The importance of producing data that agencies and scientists can rely on is clear to River Watch participants at all levels. To understand what factors affect the reliability of River Watch data, we turn to the other two lenses.

Accuracy

Accuracy in this context means how well the values reported represent the actual state of the stream. With our initial data-centric approach, accuracy initially stood out to us as the "obvious" key lens for representing and measuring data quality, and this view is common in the literature [112].

Accuracy seemingly provides an obvious metric for computation: given a known ground truth, it is relatively simple to operationalize the accuracy as the minimization of error between the measured value and the ground truth. For example, Nicholson et al. [66] computed relative accuracy by comparing the average of volunteer-collected data with that of professionals. However, there are a couple of key challenges that significantly reduce the usefulness of such an approach when evaluating the quality of individual values.

You can't measure the truth. Many of the sites monitored by River Watch teams have no one else monitoring them, so there is often no other data available to make comparisons against. When sites do have multiple monitors, they will usually visit at different times or for different reasons. For example, River Watch teams typically sample at a regular monthly interval, while professional monitors may have a specific purpose for their sampling, such as going out immediately after a rain event. In this case, one might expect River Watch data to actually be *more* representative of average stream conditions than the data collected by professionals.

This brings up a broader concern with metrics that attempt to demonstrate the accuracy of volunteer-collected data versus professionally-collected data by simply comparing them numerically. A numeric difference can be computed, but it has no directionality, unless one assumes a-priori that professional data is always more accurate. One agency staffperson, while admitting that their view was not universally shared among fellow experts, had this to say:

> "I don't believe that I can take any better sample than a student – if we're both trained exactly the same way and it's something that they are physically able to perform."

Truth is unattainable. More significantly, in order to measure the accuracy of an individual data point, it is necessary to have a clear definition of the "actual value" [114]. In the field of water quality sampling this is not as straightforward as it may seem. One coordinator noted that monitoring data consists basically of point samples in a continuum of values. Fluctuations for various parameters (such as Dissolved Oxygen) happen throughout the day and continuous monitoring is needed to have a complete picture. With measurements being taken from a relatively small cross-section of space-time, it's hard to say that any single value is completely representative of stream conditions.

Accuracy is certainly important to River Watch and its partners, but it is difficult to operationalize for the reasons given above. Perhaps because of these difficulties, agencies that rely on River Watch data have a different basis for their trust. This basis can be discussed under the lens of consistency.

Consistency

Consistency in River Watch is best understood in terms of the practices used to collect the data. In order to compare data collected at two different times or by two different people, it is critical that the data is collected in exactly "the same way, every time" so that any differences that are found are real and not due to inconsistent procedures. The goal is to minimize "variables" as much as possible, so that data can be meaningfully compared.

Follow the SOPs. Consistency in River Watch is maintained through *standard operating procedures* (SOPs) that cover everything from equipment calibration to taking the actual reading. In spite of the name, it is not uncommon for various monitoring groups to create SOPs specific to their project goals. This means that while data within each project may be consistent, it cannot be easily be compared with data from other projects. This means the data is effectively useless for any large-scale analysis.

To avoid this scenario, nearly every water quality monitor in the Minnesota portion of the Red River basin, from River Watch volunteers to paid watershed district staff, have agreed since 2001 to follow a standardized set of SOPs [86] throughout the basin. These SOPs are defined and maintained by an advisory committee with representatives from River Watch, multiple local watershed management agencies, and the MPCA.

While students are usually the ones doing the actual sampling, the teacher is always observing to make sure that everyone "stays on track" and follows the SOPs. Occasionally a coordinator will come along to observe how well SOPs are being followed, but they will try to avoid directly participating in the process. To further promote consistency, each student may be assigned a specific role on the team. One student may be assigned to run the Sonde, another will make a reading from the transparency tube, and a third will take notes. One teacher noted that "We try to let the kids do what they enjoy, I think that helps with quality."

Consistency vs. Accuracy. Consistency in River Watch is certainly emphasized in order to promote accuracy, but the concepts could theoretically come into conflict. One could imagine a particular SOP that ends up "consistently" over-estimating or under-estimating a particular parameter. This hypothetical scenario was brought up to one student, who responded:

> "It could be consistently wrong, but at least it's *all* consistently wrong I guess. I mean as long as – if someone that knows what they're doing, like a qualified person knowing 'Well, this is the steps that you need to take', I think it's safe to say that as long as you follow those steps to your best ability, and everyone does the same, that you're gonna get good test results."

We noted above how an accuracy-based approach to quality might evaluate data by comparing values collected by different teams. Instead, a consistency-focused approach requires SOPs to be followed *so that* data from different teams can be meaningfully compared.

Measuring consistency. How can consistency be operationalized in River Watch? One of the standard procedures provides a metric. Each monitoring team generally visits several sites on the same day. At least once out of every 10 visits the team is expected to take a *field duplicate* measurement immediately after their first sample. Consistency can then be measured by computing the *relative percent difference* between the sample and its duplicate. If the RPD is consistently outside a given range (30% for some parameters), monitors are encouraged to "evaluate your sampling procedures and think of ways to make improvements" [86].

River Watch participants in all roles appreciate the value of SOPs and the consistency they provide. Students and teachers alike emphasized the importance of doing everything the "same way, every time". River Watch coordinators partially base their assessment of the quality of each school's data on their knowledge of how good that school is at following SOPs. Agency staff, in turn, base their trust in River Watch data on the assurance from the coordinators that SOPs are in place.

Summary

In order for data collected through River Watch to be *reliable*, it must provide an *accurate* assessment of stream conditions. The primary way *accuracy* is enforced is through *consistency*; by following standard operating procedures. Based on our findings, it seems consistency is the most effective lens through which data quality can be operationalized in the context of River Watch and similar citizen science projects.

Significantly, official agencies trust River Watch data because of the plans in place to ensure quality, not because of quality-ensuring computations they perform on the data directly. The question that coordinators and agency staff seem to be asking is not so much "Is this data completely accurate?" (hard to answer) as "Did you follow consistent procedures when collecting this data?" (easier to answer). However, in order for agencies to use River Watch data it must be provided to them. We discuss the current process for managing and submitting data in the next section.

2.4.2 How is quality maintained?

RQ2. How is community work organized to maintain quality? What kinds of errors might occur and how are they addressed?

The end-to-end process of collecting and submitting River Watch data to the MPCA has a number of steps, many of which involve data entry or manipulation and are not explicitly covered by SOPs. We wanted to identify any potential deficiencies in the current process that could be addressed by technological means. We therefore asked participants how errors occur, how they are identified, and how they are resolved.

Potential Sources of Error

From the interviews, we learned that essentially all potential errors in the data fall into these two categories:

- **Data Collection Errors.** These are errors "at the source" due to miscalibrated equipment or incorrect methods. As noted above, these errors are broadly prevented by following standard operating procedures.
- **Data Entry Errors.** These are errors that can happen at any point after the data is collected, including errors in writing on the field sheet and errors copying data from the field sheet into the computer.

As soon as the values are measured, they are recorded on a standardized field sheet. Typically one student calls out the measured value while another records it. Potential errors include illegible handwriting, or misheard numbers. Included in the data are field notes – observations of the surrounding area, such as weather and even if there are birds around. These notes can seem to be arbitrary and unimportant, but they play a critical role for explaining anomalous values.

Later on, data is copied from field sheets into a standardized Excel spreadsheet. During this transition there is a potential for typos (such as misplaced decimal points), incorrect column usage, or incorrect site identifiers. Teachers noted that ideally data should be transcribed as soon as possible after making a trip, in case there are any questions about the field sheet that need to be addressed by the person who filled it out. However, they also noted that in practice students and teachers are busy, so this is usually only done at the end of the sampling season.

Submitting data. Once a year the coordinators merge all of the River Watch data for the year and submit it to a centralized database used by the MPCA. In years past, the coordinators did this by manually combining data from all of the schools into a single master spreadsheet. Now, teachers (or students) can upload their spreadsheet data directly into the web-based database system, from which coordinators can export a merged spreadsheet for submittal to the MPCA.

Once data is successfully submitted and accepted by the MPCA, it becomes a part of the official record. Since data considered to be of poor quality does not usually make it this far, saying that data "gets submitted" is roughly equivalent to saying it is considered to be of good quality in the parlance of River Watch participants.

Identifying Errors

Because of the potential for data entry errors and data collection errors to affect the quality of the dataset, we also asked participants what practices were in place for catching errors after they occurred.

Domain knowledge. River Watch participants and coordinators said they could often tell when a data value is likely to be suspect. This knowledge is based on experience working with the equipment and with previous data at the same site. Errors that occur while sampling can be caught by those familiar with the data; even students said they were capable of "just knowing" when a particular value is wrong and might need to be re-taken.

The coordinators also mentioned double-checking data by looking at the expected correlation between related values. For example, there are a number of subjective and objective parameters related to the clarity of the stream with a high level of correlation. If one is high and another is low, the coordinators would suspect those values.

Outlier detection. River Watch participants and coordinators discussed various technological means for catching errors in the data. A rudimentary but effective approach cited by coordinators and agency staff was to put all values into a spreadsheet and sort them - making it easy to see which values were outliers.

In addition, the web-based database system used to manage River Watch data incorporates mechanisms for validating data when it is uploaded. This process includes both rejecting values that fall outside the valid range of a parameter (such as a negative pH value), and drawing attention to values that are legal but outside the expected historical range for a particular site.

Recovering From Errors

The mechanism for addressing errors once they are found depends on whether or not they are recoverable. We learned that data entry errors are generally recoverable, while data collection errors are generally not recoverable.

Recoverable errors are usually data entry errors that happened somewhere along the way after the initial value was recorded. They can usually be fixed by updating the electronic data to match what is written in the field sheet. The web-based system supports the ability to correct values after they are submitted, though this was not always the case.

“No data > bad data?”. If the data matches the field sheet but still seems to be out of the expected range for a value, then it is possible that SOPs were not followed correctly. If there is no reasonable explanation for an outlier, the data is usually simply thrown out and not submitted - since “no data is better than bad data” - at least according to one organizer.

Of course, there may be cases where there are genuine outliers, and it is important not to throw them out. Anomalous events such as recent rainfall or upstream beaver dams can also cause outliers. In that case field notes provide some idea of what was going on. Field notes therefore play a critical role in helping the coordinators determine whether a data point is a genuine outlier or a data collection error that needs to be thrown out.

The decision of when to throw out unexplained (though theoretically valid) outliers in the data seems to be largely a judgment call, based on the coordinators’ experience working with the sites in the past. Partially due to the design of the original web system, the data is often deleted from the Excel spreadsheet before it gets uploaded, so there is currently no robust mechanism for tracking when data values are thrown out or for what reason. This means that it is not easy to measure how often this occurs or if there are lessons that could be learned from the situation.

More significantly, unrecoverable errors are a sunk cost. The automated metrics for identifying errors are valued by River Watch teams, since they can help prevent incorrect data from being relied on. However, they also can be a source of disappointment, as one teacher noted.

> "If you goof up, your data gets spit back out again by the website (thank goodness) and – it's not usable. *It's like you wasted your time.*"

The best way to prevent this from happening, this teacher said, was to "know the norm" for a site so that data collection errors would be caught when they happen, at the source rather than months later when there's "nothing you can do about it".

Summary

The quality of data can be compromised when data is collected or when it is entered. SOPs serve to prevent data collection errors, while domain knowledge and automated tools serve to identify errors that do occur. While data entry errors can be fixed by reviewing the field sheet, data collection errors usually cannot be unless they are noticed immediately. In general, the more time passes after a sample is taken, the harder it is to recover from errors. This differentiates OCS projects from open collaboration projects like Wikipedia, where article quality can theoretically be improved at any point in time.

2.4.3 How is quality balanced?

RQ3. How is quality balanced among other community goals? Can the strictness that high quality requires come into tension with those goals?

Between the rigorous enforcement of SOPs and the somewhat tedious process of entering and submitting data, it seemed possible that there might a tension between the "data collection" and "educational" goals of River Watch. We therefore wanted to determine which of the goals was more important to participants, if any, and if there was a perceived tension between the two.

More than one coordinator discussed the possibility of extending into other types of monitoring where the emphasis would be primarily on educational aspects and that "maybe some of that data would not be submitted to the powers that be." This could be important for extending River Watch to other schools less interested in data collection.

Impact of Data Quality on Education

An obvious question then, is whether or not an emphasis on rigorous data quality is seen as a burden or a distraction by those interested primarily in the educational aspects of the program. Interestingly, the teachers that we interviewed did not see any tension between education and data quality. There seemed to be two primary reasons for this.

Data is part of the lesson. River Watch teams do not merely collect data and forward it to others – they use it themselves to learn about their local environment. Every year at the annual forum, River Watch teams present posters detailing what they have found about their watershed through the data they collected. The lessons learned about the status of their streams would be effectively meaningless if the data was not correct, so teams are internally motivated to ensure quality.

Data quality *is* the lesson. Even more interestingly, the rigorous process of collecting scientific data is a core part of what teachers want to teach their students. In the words of each of the teachers:

> "I think what River Watch does is it gives us an opportunity for real-life things, and when kids get out in the real world they've got to make sure that what they do is good, it's got quality. So I think it goes hand-in-hand with the educational experience that we're trying to provide."

> "I think that you learn *by* doing things the right way, not 'in spite of'."

> "Science is supposed to be a process of discovering how the world works based upon a certain set of protocols and procedures, the scientific method . . . If you want to teach people in general the value of science, I think you have to emphasize that *how you get* those numbers is important."

Students also appreciated this aspect of the program.

> "It can help you in the long run, just making sure that when you do something you do it right."

To be sure, there are many other lessons in River Watch that teachers and students find valuable, such as an appreciation for the natural world, an awareness of environmental concerns – even just a chance to get outside and have fun. However, it was very clear for those we interviewed that data quality itself forms an integral part of the scientific education provided by River Watch.

Impact of Education on Data Quality

It seems then, that the emphasis on data quality in River Watch enhances its educational value, rather than limits it. One then wonders then if the inverse is also true - is data quality in River Watch improved by its educational focus? While more research is needed, we identified a number of potential advantages of the River Watch model.

- The educational basis of the program lends itself naturally to ongoing training. The coordinators maintain an ongoing relationship with each school to ensure equipment is maintained and SOPs are followed.

- River Watch teams span several high school grades and students generally stay in the program until they graduate. While there is constant turnover as students graduate and new students come in to replace them, older students generally (though not in every case) enjoy passing their knowledge to the newcomers.

- The funding opportunities provided through the educational environment provide an additional stability to the program that helps to ensure its longevity.

- While taking an entire class out actually tended to lower quality, teachers noted that having a small group of 3-4 motivated students with clearly defined roles allows each person to focus on their task. At least one teacher even requires students to submit an essay explaining why they want to participate before joining the team.

Summary

After talking to participants, we found that it was problematic to attempt to clearly delineate between the educational and data collection aspects of the program. As one coordinator put it, “it all kind of blends together”. We instead learned that an emphasis on data quality is itself an integral part of the River Watch educational process, rather than a competing goal. Students learn how to do good, rigorous science, and to collect and use hard data to quantify environmental issues in their region.

2.5 Discussion

River Watch is an example of a successful community-based monitoring program. By fully integrating the data collection and educational aspects of the program, it has a unique robustness that it might not have by prioritizing either aspect alone. Data quality in OCS programs like River Watch is much more than a computational attribute of the data itself – it is a qualitative process that involves a number of human factors that cannot all be easily mitigated through computational means. Nevertheless, technological interventions can support the quality assurance process by leveraging the vast amount of domain knowledge and intuition already held by participants themselves.

In the context of River Watch, *accuracy* is a key dimension of data quality that initially appeared promising. However, accuracy is difficult to operationalize due to the nature of the data. While statistical validation and absolute range limits help enforce data quality, there is a more important metric in this context: *consistency*.

Consistency in following standard operating procedures appears to be the key to understanding data quality in River Watch. Thus, I learned that the full answer to RQ1 lies in RQ2: data quality in River Watch (and indeed, many observational citizen science projects) is inherently tied to the practices and procedures for measuring and reporting the data. This led to the paper's title: *quality is a verb* [101]. With this process-centric (rather than data-centric) view of data quality in mind, I noted the potential value of process-recording mechanisms (such as wiki-style revision logs). I also noted that much of the data was not being uploaded to the repository until the end of the year, when there is little that can be done to repair data that is obviously incorrect. In Chapter 4, we discuss potential technical interventions for these challenges.

I was pleasantly surprised to find that there is not a perceived tension between educational goals and data quality practices. Instead, both teachers and students indicated that they felt that rigorous quality control was a key part of the River Watch experience. That said, there may have been a selection effect in play: I only interviewed participants from active schools, so any schools that had dropped out were not represented in my limited dataset. More broadly, I was interested in better understanding the factors that promote volunteer retention in OCS. To better understand these questions, I turned to CoCoRaHS, a much larger project.

Chapter 3

Measuring Data Quality: Volunteer Retention in CoCoRaHS

3.1 Introduction

In the last chapter, we examined how data quality and participant engagement are intertwined in observational citizen science. We focused particularly on *internal* dimensions of data quality, as reflected in the processes and procedures used to create the data. This qualitative focus was partly due to the nature of the domain, but also to the relatively small scale of the River Watch project. In order to extend and validate the work quantitatively, I began looking at a much larger project, both in geographic scale and in the number of observations. I also wanted to focus more on *external* dimensions of data quality as understood by the users of the data.

I shifted my focus to *CoCoRaHS*, or the Community Collaborative Rain, Hail, and Snow Network. CoCoRaHS volunteers submit daily observations that are used as nearly real-time input into the meteorological outputs of the U.S. National Weather Service and other organizations. While every CoCoRaHS contribution is useful, not every contribution is included in aggregate datasets like those published by the Global Historical Climatology Network (GHCN).

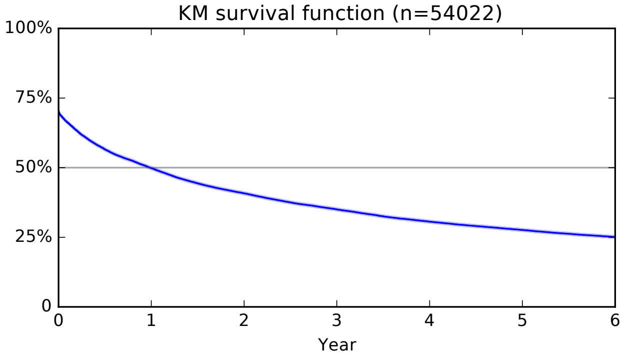


Figure 3.1: Kaplan-Meier survival curve for CoCoRaHS participants. 50% of participants drop out by the end of their first year, including 28% of accounts that never submit an observation. However, those who make it past the first year often stay for many more (c.f. [88]). Confidence intervals are indistinguishable until 10 years.

To evaluate data quality in CoCoRaHS, I began by comparing the computationally filtered GHCN dataset with the original CoCoRaHS dataset. While the rules for inclusion in GHCN include a number of statistical factors, the strongest quality signal – by far – is simply the number of contributions. Specifically, participants who have contributed 100 or more daily observations are usually included in the dataset, while those with fewer contributions are not. Thus, I found that even in a large-scale OCS project like CoCoRaHS, data quality is focused on characteristics of the data generation process, versus a computable notion of "accuracy" per se. I also found that, as in River Watch, this understanding of data quality is directly tied to questions of participant motivation and engagement.

With this in mind, this chapter focuses on *volunteer retention*, a directly computable metric that measures both the data quality and participant engagement goals of OCS. As Figure 3.1 shows, 50% of CoCoRaHS participants drop out by the end of their first year. As it turns out, this is also the ratio of CoCoRaHS participants who contribute 100 records or more, and the two metrics are highly correlated[1] . While user activity patterns are well-studied in fully online systems like Wikipedia [33] and Galaxy Zoo [60], little quantitative work to date has measured activity and retention in OCS. A better understanding of the predictors of retention would help inform recruitment strategies as

[1] 92% of those who make it past the first year also pass the 100 contribution threshold, vs. 14% of those who drop out sooner.

well as potential interventions [93]. We are interested in the following research questions:

RQ4. How well do **participant characteristics**, **task characteristics**, and **early activity** predict retention?

RQ5. How does retention relate to other measures of **data quality**?

This chapter provides three contributions toward understanding these questions.

1. We find that participant age at signup is a particularly good predictor of retention *and* (to a lesser extent) several other measures of data quality. Since very few studies have directly compared participant demographics to actual activity levels, this is the primary contribution of this chapter.
2. We find that exposure to more below-freezing days *positively* correlates with retention. This is counterintuitive given that cold weather presumably increases CoCoRaHS task difficulty.
3. Finally, we show that activity within the first month after signup is one of the strongest predictors of long term retention. While this finding largely replicates previous work in online peer production communities, it leads to important implications for the design of citizen science and crowdsourcing projects.

In the remainder of this chapter, we review related work in volunteerism and analytical methods, before turning to discuss the design of our survival analysis. We then explore the results showing how our measured characteristics correspond to volunteer retention and data quality. We discuss potential explanations for why our results contrast with earlier work and intuition, before concluding with recommendations and directions for future work.

3.2 Related Work

3.2.1 Volunteer Characteristics

While crowdsourcing and computer-aided citizen science are relatively new phenomena, they are related to much older fields of study such as volunteerism. It is known that retirees are generally less likely to volunteer than younger age groups, with volunteerism peaking at around 40-45 [76]. However, the difference in volunteering rates is much smaller in recent years than it was a few decades ago [12], and retirees who do volunteer tend to spend more actual time per month than younger volunteers [13]. There are also large differences in peak age depending on activity type, and some evidence that motivations for volunteering change as participants age [70, 71].

Several studies have explored factors that influence retention and other measures of commitment to volunteering. Participant motivations, attitudes, and beliefs are often the primary focus of research [8, 94, 95]. Cnaan et al. studied demographic, personality, and situational factors, noting that volunteerism is qualitatively different than paid work and that different theoretical models should be used [15]. They found that age was positively correlated with self-reported duration of volunteering. Komp et al. found that chronological age by itself is a poor predictor of time spent volunteering, at least for the oldest age groups [50]. By contrast, we found that numeric age was useful as a predictor in our model.

In addition to age, gender is also known to influence the types and duration of volunteer activity [76]. While not directly related to volunteerism, models for technology adoption also incorporate both age and gender. Most relevant to our work, age and gender are known to interact to influence habit forming, with the effect being strongest for older men [109]. Research on online social networks has shown that perceptions of existing gender imbalances can create a self-perpetuating skew toward one gender or the other [62].

Raddick et al. studied demographic patterns in Galaxy Zoo, finding that older participants are slightly over-represented in surveys (versus what would be expected from the general internet population) [83]. However, Cox et al. found no significant effect of age on retention or activity [18]. Kobayashi et al. found that seniors remained more active in contributing to an experimental OCR proofreading system [49]. Similarly, Baruch et al. found that the most active participants in the Tomnod platform tend to be over 50 [3], based on the results of several surveys. However, they do not report quantitative differences in the actual number of contributions or in participant retention.

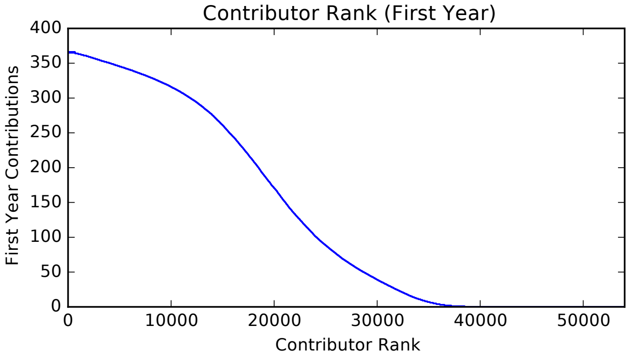


Figure 3.2: While there is a long tail of CoCoRaHS contributors, the power law for any given year is truncated due to the maximum contribution rate of once per day.

3.2.2 Analytical Methods

Traditionally, research on volunteerism has relied on self-reported surveys to determine activity levels and duration [124]. Relatively recently, studies of online systems have been able to combine theory-backed survey results with actual activity logs [27, 68]. A common finding is that small number of participants often perform the bulk of the work. Ma et. all characterized this “power law” distribution in OpenStreetMap, finding significant skew in user contribution rates as well as in the size and number of edits for each geographic element [59]. As Figure 3.2 shows, CoCoRaHS has a similar distribution, though it is truncated due to a maximum contribution rate. Panciera et al. found that the most active Wikipedia contributors distinguish themselves almost immediately after signup [75]. Yasseri et al. evaluated temporal activity patterns in OSM, noting important differences with Wikipedia [127].

Survival analysis has been used to study retention outcomes in both Wikipedia and OSM. Ortega et al. describe the use of survival analysis to characterize median dropout times in several Wikipedia languages as well as open source projects [73]. Zhang et al. further characterize the distinct survival functions for two main categories of Wikipedia editors [128]. Zhu et al. showed how survival analysis could be used to predict the survival of wiki communities themselves [129]. Dittus et al. evaluated participation in three Humanitarian OSM Team initiatives, comparing early activity with long-term retention [19]. Like our study, they leveraged actual contribution history with an inactivity buffer to measure survival, and they incorporated task difficulty and early activity as predictor variables. However, they did not explore how participant demographics affect outcomes.

Relatively little work to date has explored participant contribution and retention patterns in OCS. Sullivan et al. briefly characterize participation patterns in eBird, noting that activity peaks in May and drops during the summer [107]. Wood et al. note that 90% of contributions come from the most active 10% of eBird contributors [126]. Wiggins and He explore the use of technology and its affect on data validation practices in iNaturalist [120].

We are unaware of published research that characterizes the demographics of eBird and iNaturalist contributors, or quantifies factors that lead to increased participation and retention. Together with CoCoRaHS, we believe that these are among the largest OCS projects, making the relative lack of quantitative research particularly striking. We fill a gap in this research by applying techniques from peer production and crowdsourcing to study participation patterns in observational citizen science.

3.2.3 About CoCoRaHS

The Community Collaborative Rain, Hail, and Snow Network, or *CoCoRaHS*, is a multinational citizen science project engaging volunteers in daily precipitation monitoring. The network started in Colorado, USA in 1998 after an underpredicted flash flood, and has since expanded to all 50 US states, as well as Canada and the Bahamas [88]. As of March 2017, over 50,000 participants have contributed over 33 million daily observations.

CoCoRaHS participants are asked to empty a rain gauge daily and report the total precipitation during each 24-hour period. During winter months, participants can either melt snow to report the equivalent rain amount, or take a break and rejoin in the spring. Like many citizen science projects, CoCoRaHS tries to balance data collection with the equally important goal of public outreach toward scientific awareness.

Reges et al. describe the demographics and participation patterns in CoCoRaHS, noting a skew toward older participants in surveys and providing an overview of seasonal activity patterns [88]. However, they do not measure whether the skew is due to differences in recruitment or in retention rates, or directly measure the relationship between age and actual contribution activity. We build on their work by measuring actual activity levels, and controlling for initial signup skew when computing retention rates.

We demonstrate that *both* recruitment and retention are skewed toward older participants in CoCoRaHS, and that the relationship between age and retention is effectively monotonic.

In addition to a skew toward retirement age, the majority of CoCoRaHS participants are white and male. Given the project's dual purpose of data collection and scientific awareness, CoCoRaHS has set an explicit goal to expand the diversity of project participants [88]. This is important for expanding climate literacy among a broader segment of the population, but also for expanding geographic coverage in the dataset (since different regions have different demographic makeup). From a practical standpoint, applications for additional project funding often require a discussion of how the resources will be used to serve underrepresented populations.

Almost since its inception, CoCoRaHS has had a website[2] to support participant registration and online data entry. Since 2014, CoCoRaHS has also provided data entry apps for Android and iOS. The provenance and editing history is stored in the database with each observation, providing a unique opportunity to apply analysis techniques from crowdsourcing and open collaboration platforms to observational citizen science.

3.3 Hypotheses

While parts of this research were exploratory in nature, we drew heavily from prior work and theoretical foundations, which we present as hypotheses below.

The most active participants in Wikipedia and other large peer production systems have tended to be younger [30]. However, domain-specific citizen science projects have attracted middle-age and even older participants [83, 3], and Reges et al. [88] describe a skew toward middle and retirement age participants in CoCoRaHS survey respondents. We hypothesize that this skew will pan out in activity levels and retention as well.

> **H1: Age positively correlates with retention in observational citizen science.**

A critical component of participant retention in citizen science relates to the structure of the actual task being performed [119, 103]. In theory, participants in a citizen

[2] `https://cocorahs.org`

science project all follow the same protocol for collecting and reporting data. In practice, the experience of task completion may vary significantly for different participants depending on how often they experience the phenomena of interest. For example, in Galaxy Zoo, participants who see fewer interesting photos of galaxies are more likely to end their session early [60]. Thus, our second hypothesis is as follows.

H2: Frequent encounters with the monitored phenomena improve retention.

With field-based citizen science projects, the design of the data collection protocol is particularly important [101]. In some cases, participants may need to use a more complex protocol to adapt to changing external conditions. This may lead to increased task difficulty. We hypothesize that as task difficulty increases, so does the likelihood of participants dropping out. This could be due to exhaustion, or due to reduced participation levels. In particular, when the task difficulty changes based on geographic and seasonal factors, we would expect this to be reflected in retention patterns as well.

H3: Increased task difficulty leads to lower retention.

In many open collaboration systems, long-term participants differentiate themselves within a short period after signing up [75]. However, in Galaxy Zoo and OpenStreetMap, high initial activity is sometimes associated with early burnout [19, 77, 96]. Given the upper limit on CoCoRaHS participation of one record per day, we expect a relatively straightforward relationship between early activity and retention.

H4: Participants who are more active during their first month stay longer.

Finally, we expect that highly motivated participants will not only participate longer, but also contribute higher quality data than less active participants [2].

H5: Participants who are more active during their first month contribute higher quality data throughout their first year.

3.4 Methodology

Our primary dataset was the full archive of 35,581,914 CoCoRaHS daily observations from June 1998 to February 2017[3] . We excluded multi-day observations, which make up a relatively small fraction of the dataset and are considered less useful by meteorological analysts. Since the CoCoRaHS database incorporates data from other monitoring networks, we limited our analysis to participants marked as being members of CoCoRaHS. We also incorporated information entered in the account registration form on the CoCoRaHS website, which includes information about each participant and the geographic location of their rain gauge.

Our dependent variables were volunteer retention and data quality, as defined below. We operationalized 10 characteristics of the participant, the task, and early activity as independent variables. All independent variables were defined using only information that would be known within a month after a participant signed up.

3.4.1 Dependent Variables

Volunteer Retention

To test hypotheses 1-4, we conducted a survival analysis using the Cox proportional hazards model. Survival analysis has an advantage over other statistical methods in that it can handle "censored" dropout events (i.e. for participants who remain active past the end of the study window). In addition, the use of survival analysis allows us to separate the *outcomes* of various predictor variables from any biases in their initial distribution.

Since participants do not announce when they are leaving, we determined dropout based on the period of inactivity after the last contribution. While Dittus et al. use 180 days as a cutoff [19], Karumar et al. use 365 days [44]. Given the large seasonal variability in CoCoRaHS participation, we limited our choice of time periods to multiples of a year. We settled on a one year cutoff, which corresponds to CoCoRaHS' internal inactivity determination rule [88]. As it turns out, about 8% of participants who we counted as having dropped out rejoined again after a break of over one year. For these participants, we excluded the second activity period from our analysis.

[3] As determined by observation date, which is not always the same as data entry date.

We used the account creation date as the start date for our analysis, unless the participant submitted an observation for an earlier date (about 8.5% of accounts). We necessarily excluded any participants who signed up after February 2016. We counted participants who signed up and never contributed anything (about 28% of accounts) as having dropped out on their first day.

We conducted our exploratory analysis with Kaplan Meier curves using the `lifelines` survival analysis library for Python [22]. We used `pandas` and `matplotlib` to create the graphics in this chapter. We used R's `survival` package to run the final Cox analysis due to its support for frailty terms in the model (see the subsection on control variables).

Data Quality

To test Hypothesis 5, we also incorporated several additional metrics of data quality, building on our prior work with River Watch. In addition to the three internal metrics discussed in Chapter 2 (*accuracy*, *reliability*, and *consistency*), we introduce a fourth external metric, *timeliness*, given the particular focus CoCoRaHS has on getting observational data to the National Weather Service in near real-time. We can view participant retention as a measure of reliability, though this is somewhat different than the operationalization discussed in Chapter 2. We can operationalize the other three metrics as follows:

- **timeliness**: the proportion of observations entered in the system on the same day they were observed
- **consistency**: the proportion of observations that immediately followed a previous day observation (i.e. without multi-day gaps, which are less useful for analysis)
- **accuracy**: the proportion of observations that were never edited[4]

[4] While an edited record is likely higher quality than the unedited version, we take the act of editing as a proxy for an accuracy issue in the original data. This is based on prior work on Wikipedia which treats the persistence (i.e. non-editing) of each word as an indicator of article quality (c.f. [34]). However, note that in CoCoRaHS, editing is relatively rare and usually done by the contributors themselves. In addition, as discussed in Chapter 2, it is challenging to define a universally meaningful notion of quality and accuracy [101]. We consider this particular operationalization of accuracy to be good enough for this particular use as one of several proxies for participant data quality.

Since we are already measuring retention (reliability) separately, we designed the three other metrics as percentages to minimize the effect of the total number of contributions. However, even with this factoring, the metric for consistency is necessarily correlated with higher contribution rates. We computed these metrics for each participant's first year of contributions, excluding participants who never contributed anything. We then conducted a separate linear regression for each metric.

3.4.2 Participant Characteristics

Age

To test Hypothesis 1, we included participant age in the model. Fortunately, we were able to obtain this without a survey, as participants can optionally provide their age in years to CoCoRaHS when creating an account. However, only around 30% of participants actually provide their age. Since the Cox PH model does not support null values in predictor variables, we used multiple imputation to fill in plausible random values for age [11]. We validated our results by running a second model containing only the participants who entered age, and by comparing the actual median survival for participants who entered ages versus those who did not. As we discuss in the next section, the results for each method were compatible.

Gender

As noted previously, gender has also been shown to influence volunteerism and technology adoption. Participants are not asked for their gender when signing up for CoCoRaHS, but they do enter their full name[5] . We were able to estimate gender using the relative frequencies of participants' first names in the United States, using the equation $\frac{M(x)-F(x)}{M(x)+F(x)}$ as proposed by Liu et al. [56]. While they use U.S. Census data for their name corpus, we used birth names as registered with the U.S. Social Security service between 1950 and 2016.

We used 0.5 as the association threshold for our analysis. 86.6% of accounts met this threshold and were assigned a gender, with an average absolute score of 0.97. This means

[5] While most CoCoRaHS data is open to the public, full name and address are protected by privacy policy. We used only first names for this analysis.

that, on average, fewer than 1 / 75 people with each given name had the opposite gender than our assignment, at least based on the U.S. Social Security dataset. While this does not guarantee a correct assignment in every case (particularly for group accounts), we believe the accuracy is high enough that our results will not be affected. The 13.4% of accounts with a score less than 0.5 were assigned to the Unknown / Other group.

Table 3.1: First Name Gender Association

score	assigned gender	n	%
< -0.9	Female	12,618	23.4%
-0.9 to -0.5	Female	1,167	2.2%
-0.5 to 0.5	Unknown / Other	7,260	13.4%
0.5 to 0.9	Male	1,692	3.1%
> 0.9	Male	31,285	57.9%

3.4.3 Task Characteristics

The CoCoRaHS experience may be different for participants depending on how often they experience rain or snow. We can take the number of rain days as a measurement of **phenomena frequency**, and the number of below-freezing days as a proxy for **task difficulty**.

Rain Days (year before signup)

Participants are regularly reminded that reporting a zero is highly preferred to not reporting at all. Nevertheless, we expected to find that participants from drier climates would have more trouble with consistent data entry, per Hypothesis 2.

We spatially joined each CoCoRaHS participant's reported latitude and longitude to the PRISM climate grid of historical daily rainfall and temperature data [81]. Due to the nature of the PRISM dataset, we limited our analysis to participants with locations that were within the 48 continental U.S. states. We counted any daily grid cell with more than 0mm of precipitation as a rainy day. We used the 365 days before signup in order to avoid measuring the effect of rainy days that occurred after participants may have already dropped out. Ideally we would have included the rain they experienced

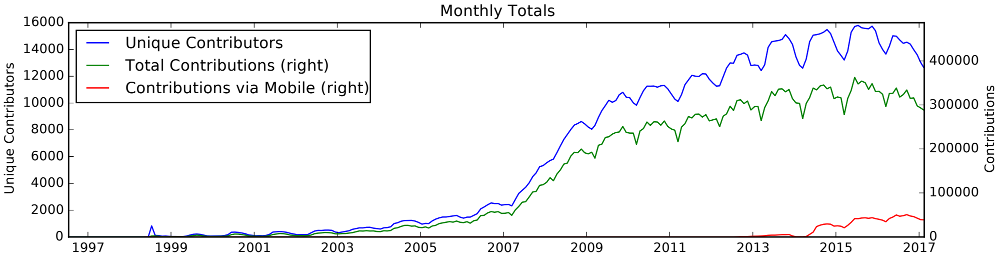


Figure 3.3: CoCoRaHS monthly contribution rates; note the pronounced seasonal effect. The right axis is scaled to about 16000x31 contributions per month

after signing up, but this is highly correlated across years.

Freezing Days (year before signup)

Per Hypothesis 3, we expected to find that participants who experienced more below-freezing days would drop out sooner, either due to the increased difficulty of the snow protocol, or to forgetting about the project during the break. As Figure 3.3 shows, there are substantial drops in activity levels during the winter months.

We calculated freezing days using the same technique as for rain days, but instead counted days where the average temperature was less than zero degrees Celsius. Given that precipitation itself changes when temperatures are below zero, we included an interaction term combining rain and freezing days.

3.4.4 Early Activity

First Month Observations

To test Hypothesis 4, we counted the number of observations each participant submitted during their first 30 days of activity. While previous work has used sessions measured in minutes or hours as a metric for activity, a timespan of a month made more sense for CoCoRaHS given the maximum contribution rate of once per day. Due to differences in signup time, some participants were able to submit 31 records during their first 30 days; we counted these participants as having submitted 30. We included an interaction term combining age and first month observations, since habit-forming is known to be stronger for older adults [109].

3.4.5 Control Variables

U.S. State (Fixed/Frailty)

CoCoRaHS is organized into separate leadership structures for each U.S. state. Each state joined at different times, with Colorado starting in 1998 and Minnesota joining in 2009. In addition, each local organization is given considerable flexibility in recruiting and participation methods, particularly during the annual March Madness recruitment drives. Some state coordinators offer free rain gauges, others create YouTube videos, while others put out press releases or letters to the editor. Some coordinators are paid to work with CoCoRaHS as part of their official responsibilities, while others are purely volunteer based. Finally, the level of enthusiasm for project recruitment differs between coordinators.

To reduce the risk of misattributing differences to climate that were instead due to differences in the local CoCoRaHS leadership structure, we included participants' U.S. state of residence as a *frailty* (fixed effect) term in the model. This may have overcorrected for effects that were in fact due to climate differences.

Timing & Other Factors

We also included a few variables to control for various factors related to the timing and nature of participant signup and early activity. We controlled for an early adopter effect by including participant signup rank within their state. We also flagged March signups since they are likely to be part of an annual recruitment drive (March Madness). We accounted for participants who had a rain gauge when signing up[6] , since we expected them to be more intrinsically motivated to start reporting right away. We also included a variable for daily internet access, since participants without it presumably need to call to submit their observations. Finally, we measured relative use of the CoCoRaHS observer mobile app for first month data entry. The apps were only available for two years of our study period, and the effects were nuanced. We hope to investigate the effects of mobile usage more fully in future work.

[6] Technically, this flag can also be set after signup, if the contributor interacts with a coordinator who has access to update their account.

3.4.6 Correlation

To ensure accurate model specification, we verified that none of our independent variables were strongly correlated by evaluating Spearman’s for each pair of coefficients. The only correlation stronger than 0.15 was between *Signup rank within state (log)* and *Rain days (year before signup)*, which are slightly negatively correlated (-0.21).

3.5 Results

The results of the Cox proportional hazards model are shown in Table 3.2. For the purpose of evaluating effect size, we compare the median survival day, i.e., the day by which 50% of participants have dropped out. For continuous variables, we compare the mean value (shown at right) and the median baseline survival (295 days) with the effect of an increase of one standard deviation in the predictor. For logistic variables, the effect size shows the difference between the false and true conditions.

We also generated figures 3.4-3.8 to further explore the effect of certain predictors on measured retention rates. We computed these by generating small bins for each continuous variable and then extracting the median dropout date and 95% confidence intervals from the Kaplan-Meier curve for each bin. While this approach loses information about the actual shape of each curve, it makes it feasible to interpret differential outcomes for continuous variables.

To provide insight into the differences between initial and long-term skews in the data, we also plot the initial distribution (as measured at or shortly after account creation) as a histogram under the retention chart for each figure. Since this is essentially the `n` for each measurement, it is inversely correlated with the confidence interval for the median outcome shown on the upper chart. There is no necessary correlation between the initial `n` and the outcome, other than potential shared underlying mechanisms.

Table 3.2: Volunteer Retention - Survival Analysis (n=52154)

Predictor	hazard ratio		95% CI	effect size	mean	stdev
First Month Observations	0.488	****	0.482 - 0.495	+1334 days	10.6 obs.	11.28
% Submitted via Mobile App	0.936	***	0.920 - 0.953	+52 days	1.6%	11.6%
Age (imputed, see discussion)	0.867	****	0.844 - 0.892	+117 days	48.3 yrs	10.44
Rain Days (year before signup)	1.027	**	1.007 - 1.047	-24 days	71.6 days	26.4
Freezing Days (“ ”)	0.948	***	0.922 - 0.975	+43 days	99.1 days	57.1
Signup rank within State (log)	1.082	****	1.065 - 1.099	-66 days	6.37	1.31
Had gauge at signup (or later)	0.662	****	0.636 - 0.690	+441 days	8.2%	
Daily internet access	0.961	**	0.938 - 0.986	+34 days	75.3%	
Signup during March Madness	1.043	**	1.012 - 1.073	-36 days	14.5%	
Female (guess from name)	1.009		0.975 - 1.043		25.5%	
Male (guess from name)	0.928	***	0.900 - 0.956	+61 days	61.3%	
First Month Obs. × % via Mobile	1.020	*	1.002 - 1.040	-18 days		
Age × Female	0.964	*	0.932 - 0.998	+27 days		
Age × Male	0.973		0.944 - 1.002			
Age × First Month Obs	0.972	***	0.960 - 0.984	+21 days		
Rain Days × Freezing Days	0.992		0.977 - 1.007			
U.S. State	(frailty)	****				
Concordance: *0.772*						
Median Retention: *295 days*						

p-values: *<0.05 **<0.01 ***<0.001 ****<2e-16

3.5.1 Participant Characteristics

CoCoRaHS is heavily skewed toward older participants: the average age at signup is 48, while the median is 52 and the mode is 60. Compared to all other age groups, participants aged 60-70 are more likely to sign up - and even more likely to stay for several years. As Table 3.2 shows, the model effect size is quite large: an additional 10 years of age corresponds to 117 additional days of participation in the program. Since the age used in Table 3.2 contains a large number of imputed values, we also ran a second model with only the participants who entered an age. The results are shown in Table 3.3. In the smaller model, the hazard ratio for older participants is even smaller and the effect is larger (+247 days).

Table 3.3: Volunteer Retention for known age (n=16196)

Predictor	hazard ratio		effect
First Month Observations	0.4710	****	+1164d
Age	0.7578	****	+247d
Female (guess from name)	0.9384	*	+41d
Male (guess from name)	0.8720	***	+105d
Age × Female	0.9561		
Age × Male	0.9614		
Age × First Month Obs	0.9559	***	+40d
Concordance: 0.781			
Median Retention: 278 days			

This striking result can be explored further by examining the actual median survival for different ages, as shown in Figure 3.4. Between the ages of 20 and 70, and in particular between 45 and 65, there is an almost perfectly monotonic relationship between age and retention. The median dropout for participants aged 19-20 is 13 days, while the median dropout for ages 69-70 is 3.12 years. Thus, Hypothesis 1 is strongly confirmed. Interestingly, the age with the highest retention (65-66) appears to be older than the peak signup age (59-60).

We also computed the median survival for participants who did not enter an age, to verify that the results could be generalized between the groups. If age was truly missing

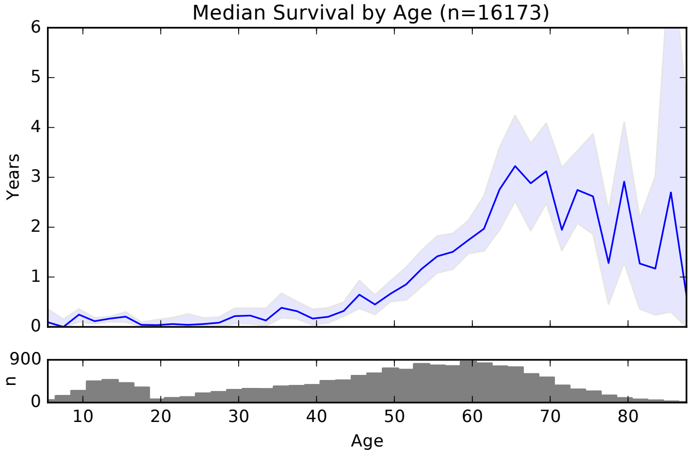


Figure 3.4: Retirement age participants are the largest group and participate for the longest. (Bin size=2 years)

at random, we would expect the median survival for unknown age to be roughly equal to the survival for all known ages (253 days). Instead, we found that the median survival for participants who did not provide an age is 405 days, which is close to the survival for participants aged 53-54. A plausible explanation is that older participants are less comfortable providing their age, and are thus are even more over-represented in CoCoRaHS participation than these results show.

In Table 3.2, each identified gender is separately contrasted with accounts for whom a gender could not be automatically determined (e.g. group accounts). Compared to women and other accounts, men are much more likely to sign up for CoCoRaHS, and participate for 61 days longer (according to the model). The actual median survival for men is 444 days, while the median survival for women is 207 days.

While a full analysis of motivations for participation in CoCoRaHS is beyond the scope of this chapter, one key question is whether the skews in age and gender are due to inherent interest in the task, or due to biases in the recruitment process (c.f. [62]). Upon account creation, CoCoRaHS participants are provided an opportunity to enter the method by which they were referred to the project. We analyzed this free-form text to determine which words were used most often.

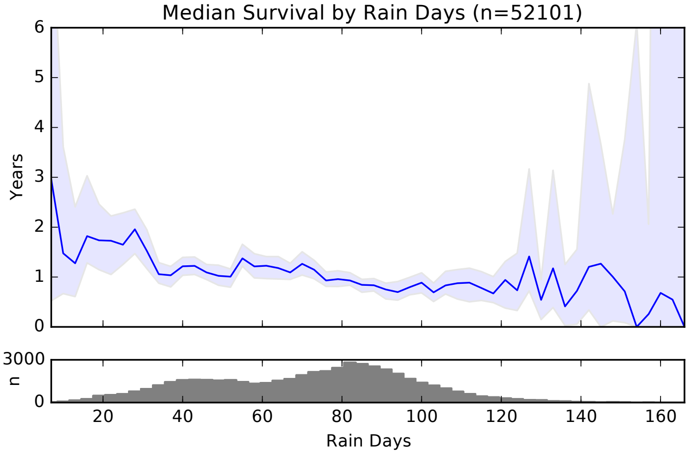


Figure 3.5: Participants who experience more rain drop out a bit sooner. (Bin size=3 rainy days)

Out of the 37,438 participants who entered referral information, the five most common words used were “NWS” (12.0%), “weather” (8.6%), “newspaper” (8.6%), “friend” (5.5%), and “NOAA” (4.9%). NOAA and NWS both refer to the U.S. National Weather Service, as do most instances of “weather”. While not conclusive, this corresponds with CoCoRaHS’ internal estimation that a large subset of active CoCoRaHS participants find the project through interaction with National Weather Service programs, such as in-person severe weather training or online weather analysis tools. Thus, it is likely that the largest driver behind CoCoRaHS participation is inherent interest in the domain and/or task, rather than relationships with existing contributors[7] .

3.5.2 Task Characteristics

Contrary to our expectations, increased precipitation appears to be correlated with a slight decrease in retention. According to the model results, an additional 26 days of rain during the year before signup corresponds to a median dropout date 24 days earlier. This relationship bears out in Figure 3.5, which shows that the actual effect is even stronger when not adjusting for state differences. The median dropout for participants

[7] Certainly, there may be demographic biases in NWS program participation that then transfer to CoCoRaHS. For what it’s worth, CoCoRaHS regional coordinators (many of whom are NWS staff) are younger, but more likely to be male, than the average CoCoRaHS contributor.

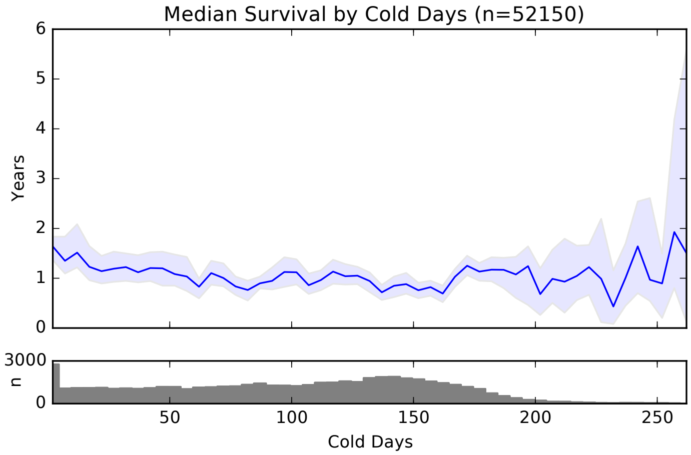


Figure 3.6: The upward trend for cold days is less apparent in this chart, which does not account for state organizational differences. (Bin size=5 freezing days)

who experienced 60-62 rainy days is 448 days, while the median dropout for participants who experienced 90-92 rainy days is 274 days. This contradicts Hypothesis 2, which proposed that more exposure to rain would improve retention.

Also contrary to our expectations, more cold weather does not appear to negatively correlate with retention. According to the model, 57 additional freezing days corresponds to a moderate *increase* in retention of 43 days. Interestingly, when running the model without controlling for U.S. state, the effect for freezing days is to *decrease* retention by 17 days, which is more in line with Hypothesis 3.

To further understand this discrepancy, we wanted to explore if the state-normalized result was being skewed by one or two geographically unique states that just happened to have more registered CoCoRaHS accounts. To check this, we analyzed the effect of freezing days on retention outcomes in each of the 48 continental states. Within each state, we split participants into those who had more or less than the average freezing days for all participants in that state. For the sake of completeness, we performed the same calculation for rain. We used a logrank test to compare the statistical significance between each pair of populations.

Table 3.4: Between-State Differences (n=51969)

Statistical Effect	States	% Volunteers
Cold ⇑ Retention	CO,TX,NC,KS,NY,OH,NJ,NH,VT	37%
Cold ⇓ Retention	MN,WY	4%
Cold Not Signif.	37 states	59%
Rain ⇓ Retention	NC,FL,TN,MO,WA,GA,AL,IA	22%
Rain ⇑ Retention	NY,MN,AZ,ME,NH,VT	9%
Rain Not Signif.	34 states	69%

The results are shown in Table 3.4. We found that in 9 states (representing 37% of CoCoRaHS participants[8]), participants in colder areas remain active longer. These include Colorado and Texas, the two most active CoCoRaHS states with over 5,000 registered accounts each. Only 2 states see an opposite trend: Minnesota and Wyoming. As it turns out, these two are among the coldest states in the U.S - though New Hampshire and Vermont are almost as cold. While more work is needed, it would appear that the positive effect of cold on retention only holds to a point. Still, the plurality of evidence points to freezing days being an indicator of improved retention, disconfirming Hypothesis 3.

3.5.3 Early Activity

CoCoRaHS participants submit an average of 10.6 records during their first month. As Figure 3.7 shows, the actual distribution is bimodal: the two largest groups are those who submit nothing during their first month (38.9%), followed by those who submit every day (5.2%). As is common with many collaborative projects, this early activity is a very strong predictor of eventual longevity. According to the model, an additional 11 contributions during the first month corresponds to an additional 4 years of participation in the program. In actuality, participants who do not contribute during their first month often never do, while the median survival for those who submitted reports for all 30 days is 5.4 years. Thus, Hypothesis 4 is confirmed.

[8] The percentage of volunteers that live in one of the listed states. Note that this does not necessarily correspond to the actual population of each state.

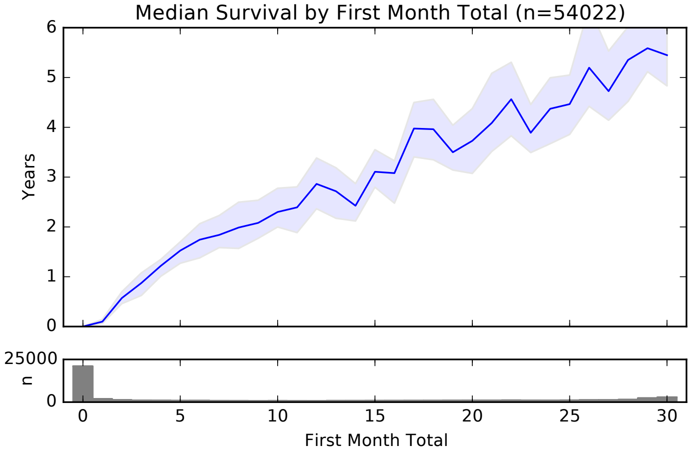


Figure 3.7: Initial activity is highly predictive of retention. (Bin size=1 daily contribution)

3.5.4 Data Quality

We also measured the effect of each predictor on the three supplemental data quality metrics, to contrast with the primary metric of reliability measured as retention. The full results are listed in Table 3.5. Note that the goal of this research is to test the effect of the predictors on several alternative formulations of data quality, not to find the best fitting model per se[9] . Also, note that the coefficients for the three models are not directly comparable. For example, the effective range of the accuracy metric is about one fifth that of consistency and timeliness, since fewer than 4% of records are ever edited.

[9] As discussed previously, the total number of contributions is a particularly strong signal for data quality in this domain. Since the number of contributions is directly tied to retention, the survival analysis itself is likely already the best model for measuring data quality. We defined the other three metrics as ratios to factor out the total number of contributions, to see if the primary results would hold up under alternative operationalizations of quality. Even with this factoring, participants who contribute more records overall will still necessarily score higher on the consistency metric, since there will be fewer gaps between submissions. Thus, the consistency model has the best fit, since the total contribution signal shows up both in the predictors and (indirectly) in the dependent variable.

Table 3.5: Data Quality Linear Models (n=37491)

Predictor	Timeliness		Consistency		Accuracy	
First Month Observations	-2.39	****	16.16	****	0.93	****
% Submitted via Mobile App	1.41	****	-0.62	***	0.30	***
Age	1.48	***	3.08	****	0.78	***
Rain Days (year before signup)	1.58	***	0.45	*	0.13	
Freezing Days (“ ”)	-0.31		-0.09		0.28	**
Signup rank within State (log)	2.29	****	-2.11	****	-0.45	***
Had gauge at signup (or later)	-1.65	***	3.54	****	-0.16	
Daily internet access	2.62	***	0.80	**	-0.09	
Signup during March Madness	0.83	*	2.39	***	0.44	***
Female (guess from name)	0.55		2.45	***	0.47	**
Male (guess from name)	2.37	***	0.82	*	0.17	
First Month Obs. × % via Mobile	-0.40	**	0.46	***	-0.08	
Age × Female	0.16		-0.31		-0.40	**
Age × Male	0.07		-0.32		-0.25	
Age × First Month Obs.	1.50	***	-1.71	***	-0.21	***
Rain Days Freezing Days	-0.10		0.47	**	0.07	
U.S. State	(fixed)	****	(fixed)	****	(fixed)	****
Linear model fit (Adjusted R^2)	*0.02*		*0.32*		*0.02*	
Mean score for metric	*69.9%*		*69.3%*		*96.9%*	

p-values: *<0.05 **<0.01 ***<0.001 ****<2e-16

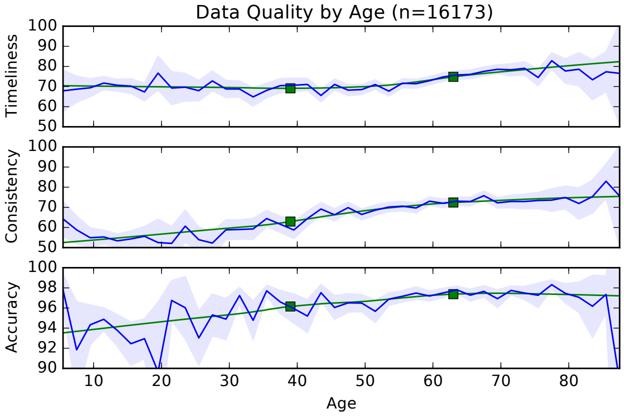


Figure 3.8: Older participants are more timely, consistent, and accurate in their data entry, though the effect is small. The smoothed line is a Lowess curve with zero iterations, with squares for the 25th and 75th percentiles used to compute Cohen's d.

Interestingly, age positively correlates with retention / reliability *and* all three of the other quality metrics, though the effects for the latter were relatively small. Older participants are not only more reliable, but are also more likely to enter their data in a timely, consistent, and accurate manner. These effects are further explored in Figure 3.8. We calculated effect size by computing Cohen's d as $\frac{m2-m1}{S}$ where $m1$ and $m2$ the smoothed average scores for participants aged 39 and 63 (representing the 25th and 75th percentile ages, respectively) and S is the standard deviation of all scores. The effect sizes for consistency (d=0.32) and timeliness (d=0.21) are small while the effect for accuracy (d=0.13) is very small (c.f [98]).

The effects for other predictors are not as consistent as those for age. While men are somewhat more reliable and timely, women are more consistent and accurate in their reports. This illuminates the importance of understanding the multifaceted nature of data quality. There were generally limited interaction effects between age and gender, other than minimal evidence that older women are slightly more reliable and slightly less accurate than would be expected from their age and gender separately.

While participants who experience more rain drop out slightly sooner, they are more timely in their reports (d=0.31 for participants with 89 versus 50 days of rain). It is likely that rain prompts a more immediate response, while those who get no rain may consistently report the 0 but not in a timely manner.

As expected, there is a high correlation between the number of records entered during the first month and overall consistency for the first year. The effect size is large (d=1.13 for participants with 25 versus 3 submissions). However, timeliness is somewhat decreased for the highest contributors (d=-0.12). Thus, Hypothesis 5 is not fully confirmed.

3.6 Discussion

3.6.1 Participant Characteristics

There is a definite skew toward older participants in CoCoRaHS, not only in terms of initial sign-up rates, but also in terms of retention and (to a lesser extent) other measures of data quality. This is in line with our first hypothesis. While this finding may not be surprising to those familiar with CoCoRaHS and similar observational citizen science programs, it is relatively unheard of in the broader domains of crowdsourcing and peer production. For example, Wikipedia is heavily skewed toward younger contributors, with half of survey respondents being younger than 22 [30]. Within volunteerism research, the general consensus is that that volunteer activity peaks at age 45 [76].

With this in mind, what makes CoCoRaHS different than Wikipedia and many other forms of volunteerism? We propose three key mechanisms that may be responsible for this finding:

- First, CoCoRaHS is designed to be incorporated as a daily routine, and the strength of habit-forming is known to increase with age [109]. Several elderly participants (and their spouses) have sent messages to CoCoRaHS thanking the project for providing a stable routine and a reason to get up early every morning (c.f. [8]). Many elderly CoCoRaHS participants remain active until they are physically unable to check their gauge.

- Second, CoCoRaHS is structured around repeated monitoring at a single location, usually in the participant's own backyard. From a purely practical standpoint, this means it is more accessible to individuals with a backyard, and those who spend more time at home.

- Third, the task itself is relatively intuitive and structured around a topic of immediate general interest and applicability - local rainfall.

We suggest that further quantitative research is needed to measure the effect of different task designs on demographic interest in participation in peer production systems and citizen science. With regard to age distribution, Wikipedia and CoCoRaHS appear to be opposite ends of a spectrum, while Galaxy Zoo appears to be near the middle (c.f. [18]). We predict that the peak retention age in eBird is older than Galaxy Zoo, but younger than CoCoRaHS, given that birding often requires travel.

While increased diversity is a goal, CoCoRaHS' skew toward older participants could also be seen as an opportunity - e.g., to promote broader scientific literacy among an influential demographic.

3.6.2 Task Structure

More rainy days correspond with slightly decreased retention, while more freezing days correspond with a moderate increase, meaning Hypotheses 2 and 3 were not supported. The result for freezing days is particularly counterintuitive, as we intended it to be a measure of task complexity. As Dittus et al. and others have found, task complexity is generally known to negatively affect retention [19]. One plausible explanation is that the complexity of the CoCoRaHS task is well known and relatively stable. Participants from moderately cold locations who decide to join CoCoRaHS are already likely highly motivated and self-selecting, which would correspond to our findings related to early activity. It would be helpful for future work to quantify the geographic disparity in CoCoRaHS signups - both in terms of population and in terms of climatology.

3.6.3 Early Activity

First month contributions are this strongest predictor of long-term retention, which confirms Hypothesis 4 and replicates prior work on Wikipedia [75]. A key follow-up question, then, is whether early interactions with a project extrinsically influence retention, or whether both early activity and retention merely reflect existing intrinsic participant motivation. In Wikipedia, at least, the latter, "intrinsic" view seems to be more applicable [75].

Indeed, the majority of CoCoRaHS participants appear to join with an existing interest in the project, and early intervention efforts to date have shown mixed results. In addition, CoCoRaHS staff report encountering certain "personality types" that simply enjoy the daily routine and data management aspects of the project - sometimes even without a particular interest in the weather. With this in mind, one potential application of this research might be to recruit a wide range of participants and then focus follow up efforts on the highest-contributing participants.

However, this approach is complicated by the importance of diversity and inclusivity to the goals of CoCoRaHS and other citizen science projects. If project resources are devoted only to engaging the most active contributors, existing biases may become more entrenched and opportunities for growth may be missed (c.f. [62]). Broader recruitment activity can help expand the project's demographic base, but it may be just as important to understand why certain participants leave, as why they are less likely to sign up. In addition, research on Wikipedia has shown that there is value in using targeted interventions to promote the retention of newcomers [33].

The requirement of a rain gauge is a potential barrier to entry in CoCoRaHS, which is why many citizen science projects do not require any equipment at all [103]. On the other hand, CoCoRaHS staff note that asking participants to spend at least the cost of shipping on a gauge can help participants gain a sense of commitment that they might not have if the gauge is given for free.

It is still possible to promote early data entry while waiting for equipment to arrive. CoCoRaHS staff occasionally encourage participants to submit their initial record on a dry day (since the value is sure to be zero). However, there is no place in the CoCoRaHS database for rain measurements made without an official gauge. As an alternative, CoCoRaHS has an informal relationship with mPING, an independent mobile app for reporting precipitation that does not require an established site or any custom equipment [20]. Participants unable to commit to the full CoCoRaHS protocol can start by contributing to mPING instead.

We suggest that these types of partnerships provide a rich opportunity to engage a broad range of participants in the projects that best suit their interests and abilities. In addition, centralized volunteer recruitment platforms like SciStarter [38] can facilitate these relationships while also tracking participant demographics and activity across multiple citizen science projects.

3.6.4 Data Quality

While age consistently correlated with higher scores on all quality metrics, first month activity did not, so Hypothesis 5 was not fully supported. This may in part be due to the challenges in operationalizing data quality; our efforts to factor out the strongest signal (number of contributions) caused the remaining metrics to be quite noisy. Nevertheless, some interesting patterns arose from the data. Participants who contributed more data were more consistent but less timely than other participants. There is likely a moderate subset of users who consistently make observations every day but only occasionally enter them in the website *en masse*. This would confirm previous work noting that data entry is not considered a desirable task by many citizen science participants [103].

3.6.5 Limitations

This project focused only on predictors that could be determined within one month of signup. This simplified the model at the expense of some practical applicability. In future work, it would be valuable to use time-dependent covariates: e.g. how likely are you to drop out this month, given the number of rainy days last month? It would also be informative to examine the 10% or so of participants who did not contribute anything during their first month, but later became active (perhaps after obtaining a rain gauge). Further, while median retention was useful as a comparative metric, it also masked a large variability in individual outcomes.

This project focused only on a descriptive analysis of existing data, and our hypotheses were not fully formed prior to starting the exploratory analysis. A next step to continue this work would be to experimentally evaluate the effect of one or more interventions on retention and data quality. Another next step would be to interview a number of CoCoRaHS participants to shed more light on the differences we found in activity levels.

To simplify our analysis, we assumed that each CoCoRaHS reporting station was run by a single observer, accounting for group accounts only in our treatment of gender. In fact, a small but sizeable subset of accounts belong to school teams and other groups.

Future work could examine other indicators of group accounts and evaluate differences between groups and individuals. In addition, we did not account for the possibility of participants remaining active after moving to another house, which would require establishing a new monitoring site and account.

Finally, we focused on factors that predict individual outcomes, without fully measuring the effect of organizational structure on retention. We included a term for U.S. state, but only as a fixed effect. In future work, it would be valuable to quantitatively measure the effect of specific organizational factors and recruitment strategies.

3.6.6 Conclusion

As is often the case for citizen science, the implications of this study depend on the goals of each particular project. Some predictors of retention are amenable to intervention, while others are not. But more fundamentally, there is a potential tension between different intervention strategies. If robust data collection is prioritized, a project might focus on recruiting and retaining participants who demonstrate an ability to remain active long term. On the other hand, if educational and inclusion goals are prioritized, projects might focus recruitment and intervention strategies toward students and underrepresented groups. Projects that focus on both goals (like CoCoRaHS) will need to carefully weigh the benefits of each approach.

While crowdsourcing and peer production are often dominated by younger contributors, OCS projects like CoCoRaHS appear to have very different demographic characteristics. This is likely due to the unique structure of the data collection workflow in OCS. In the next chapter, we generalize and elaborate on this workflow and draw implications for the design of systems for improving data quality in these projects.

Chapter 4

Improving Data Quality: Workflow and Provenance Models for OCS

4.1 Introduction

As discussed in the previous chapters, there are a number of mechanisms for measuring and ensuring data quality in observational citizen science. VCS projects can often verify results by having multiple volunteers complete the same task. However, in OCS, the data are much more individualized, containing time and location-sensitive information. These data are often unique observations of changing natural phenomena and may not be directly verifiable. The distributed nature of participation means that supervision is often minimal and organizers must trust volunteers to follow instructions accurately. Thus, there is a strong emphasis on precise task definition and training to reduce error and maintain the comparability of data points, and expert review to validate the data.

Expert review is perhaps the most popular and broadly used mechanism for data validation in OCS projects [121]. The experts conducting data quality review may be volunteers, staff, or affiliated scientists who evaluate incoming data. Among other approaches, data review may take the form of a project leader reviewing data summaries for outliers (e.g., The Great Sunflower Project), an intern entering hard copy data sheets into a database (e.g., Mountain Watch), or a global network of expert volunteer data reviewers using an integrated, customized, distributed data review tool (e.g., eBird).

One might assume that observational data remain unchanged as long as they are recorded as intended by the original observer. In practice, however, *some data do change*, usually through the review process. A common procedure is checking for and flagging or removing outliers. This is not as simple as just marking a record as “invalid”, because the initial data review may not be the final word. For example, in some data sets that include species for whom range shifts are observed, outliers might be judged invalid upon initial review. If additional evidence of a range shift accrues over time, however, the original records would be re-reviewed and marked as valid. Where supported, data may also be changed by the individuals who submit it, e.g., due to post hoc development of further expertise, or after communication with a reviewer clarifies an uncertainty.

In addition to tracking changes, it is important to explicitly track the review status of data. Has it been reviewed and approved by the coordinators? Has it been submitted to and accepted by official third parties (like the MPCA in the case of River Watch)? Tracking this information would allow OCS platforms to provide options for using all data or only data that has been fully verified and submitted, and would also encourage participants to upload data as soon as possible, without needing to worry about people relying on it before it has been reviewed.

Finally, the task definition (protocol) itself can change, or multiple protocols may be supported; knowing which protocol and which “version” of the task was in place is necessary for accurate data interpretation. This is particularly important during the early stages of project development, as establishing workable procedures and quality control mechanisms can take several iterations, each of which may require a full field season.

With this in mind, it is clear that projects should ideally track changes to data for a number of reasons:

- for accurate provenance;
- in case future reversal is needed;
- to facilitate process improvement;
- to support participant skill development; and

- to demonstrate scientific rigor through appropriate data documentation.

A number of recent developments in ICT for citizen science have led to promising approaches for managing data. However, the complexity of the data management task means that familiar general-purpose tools like Excel may still be preferred by participants. Even if a project has the resources to build or contract custom ICT for data management, important revisions to the data can and do continue to occur externally.

In this chapter, we explore a *general workflow* common to many field-based monitoring projects. We discuss five citizen science projects as examples and identify aspects of data management important to each step of the workflow process. We then present a *new data model* for field monitoring workflows that handles some of the complexities inherent in managing these kinds of data.

4.2 Related Work

Technologies supporting public participation in scientific research are swiftly advancing, but with little consideration for the complexities of data management processes. To date, general-purpose ICT rarely support quality control processes such as data review outside of simple moderator approval. Metadata standards focus on documenting data rather than facilitating retention of provenance, and while wikis might seem a suitable solution to some key concerns, most citizen science projects violate the assumptions upon which wikis are built.

4.2.1 General-purpose ICT for citizen science

Despite the centrality of data in these projects, scientific data management is not a skill that project coordinators necessarily bring to the table, nor do partnering scientists or participants. Therefore, guidelines for data management plans, policies, and practices for these projects have only recently begun emerging [117, 57]. Persistent challenges in storing project data and metadata include suboptimal ICT for smaller projects [118]. Cumulative data sets, such as those produced by ongoing, long-term projects, also defy some of the usual practices in data management because there is no “finished” data product to archive as a standalone object.

Given these challenges, several tools have been developed to make it easier for new projects to get started. Online systems like CitSci.org [64], Sensr [46], and CrowdMap by Ushahidi [69] provide platforms to launch new projects with little or no programming knowledge. However, these systems are also harder to customize for more complex workflows; project leaders must either adapt their workflow to fit the software, pay someone to make modifications to meet their needs, or both.

Other projects like Open Data Kit [35] and wq [100] offer customizable, open source software platforms for field data collection. These platforms provide an alternative to both "from scratch" development and hosted solutions. wq's modular codebase is designed to support project-specific customization, and the framework did not previously support the complex data management workflow common to many field-based projects.

4.2.2 Provenance Models

There have been a number of efforts to create and standardize models for describing the provenance of citizen science and other types of monitoring data. Some prominent examples include Ecological Metadata Language (EML) [25], the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata [24], and the Darwin Core [116] - as well as a number of other field-specific metadata standards. More general provenance models for the semantic web include the PROV standard by the W3C [28], and the conceptual W7 model [85], designed to address the "7 W's" of provenance: what, how, when, where, who, which, and why.

These models can be useful for describing the provenance of monitoring data sets, but they do not fully address the data management challenges we explore in this work. At a superficial level, the goals are slightly different: these models standardize ways to *document* provenance, while the model proposed in this work is intended to facilitate *recording and retaining* provenance data throughout the project workflow. In this respect, the model proposed in this work has more in common with FilteredPush [115], as a platform for tracking and incorporating revisions from third parties into scientific datasets.

However, there is a more serious underlying concern. We have observed that in practice, not every party involved in a data exchange can be expected to utilize the recommended metadata standards and software tools. Thus, the model proposed in this work is intended to facilitate the incorporation of revisions to data received from multiple parties, *whether or not* those parties are actively tracking and reporting those revisions internally.

4.2.3 Wikis

Wikis are a well studied method for maintaining online repositories of community knowledge, with built-in mechanisms for versioning and some provenance tracking. Wikis generally incorporate a timeline or revision history, tracking changes to data with timestamps and usernames. Importantly, old values are preserved and changes can be reverted, supplemented by log messages to explain why data were changed. Previous research has explored the usefulness of revision logs for provenance [55, 72], and as a way to evaluate the quality control process [33, 106]. Other efforts work to extend wikis with additional provenance tracking features, for example the Shortipedia project [111]. The geowiki model [80] extends wikis with concepts useful for some citizen science use cases, including the ability to track changes to interdependent objects and fine-tune permissions on a per-user basis.

With these affordances, it may appear that wikis would be ideal for maintaining repositories of citizen science data. However, there are a number of assumptions that limit the usefulness of the wiki model for this case. First, most wiki models assume that once data are in the system, all further editing will happen within the system, so there is generally limited support for offline editing and reintegration. As we will demonstrate, many important review tasks for field monitoring data are accomplished using external systems.

Second, the basic wiki model generally assumes the artifacts being described are relatively stable, and can be objectively described in increasingly better detail through iteration by additional contributors. As noted above, field data collection results in individualized, first-hand observations of changing natural phenomena.

Wikis generally have poor support for time series data because the main timeline most wikis track is their own revision history. Handling time-sensitive data effectively is a general challenge given the multiple potential interpretations of time and history [92].

Wiki rules and norms are interactively worked out within the wiki and versioned like other wiki objects. Citizen science protocols, by contrast, are well-defined in order to support scientific rigor, with the primary data collection tasks usually occurring offline. These procedures change only when necessary, and with careful consideration. The quality of an individual observation must be evaluated according to the task definition in place at the time of observation, as opposed to the task definition currently in place (which may differ). Data entry is typically a secondary task, separate from data collection, sampling, and other field-based tasks, which can reduce the rates of data submission. Finally, wikis frequently are not adequately usable for project participants, many of whom struggle with fairly simple UI-driven sites [118].

4.3 The OCS Workflow

In this section, we explore aspects of the data management workflow common to many field-based monitoring projects. We give examples of each step in the process and draw general conclusions about ICT requirements for these projects.

Table 4.1: OCS Projects Studied

Project	Focus	Geographic Range	Organizational Sector
CoCoRaHS	Precipitation	U.S.A. & Canada	NGO
eBird	Birds	Global	NGO partnership
Great Sunflower Project	Bees	North America	Academic
Mountain Watch	Alpine plants	New Hampshire	NGO
RRB River Watch	Water quality	Red River Basin	NGO partnership

4.3.1 Methodology

The empirical data informing the inductive model development presented here were collected for prior studies, combining my analyses of River Watch [101] and CoCoRaHS [102][1] with Andrea Wiggins' research on three similar projects [118]. These independently conducted studies employed qualitative field research methods and shared a primary focus on *data management processes* and *ICT infrastructures* in citizen science

[1] Note that the CoCoRaHS analysis was still in early stages when this paper was written. This thesis is organized thematically rather than chronologically.

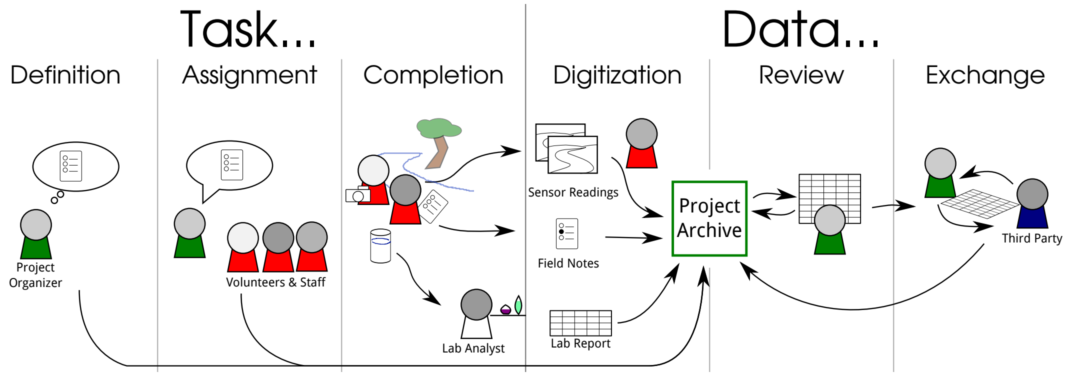


Figure 4.1: The Observational Citizen Science Workflow

projects. Data sources included interviews with project organizers, internal documents such as database schemata, and longitudinal participation and observation documented via field notes. Each study included standard procedures for ensuring research quality; for example, project leaders were invited to review and correct the penultimate drafts of analyses.

As such, the data sources and thematic focus of these studies were highly compatible. Further, the combined cases are reasonably representative of the known population of North American citizen science projects that focus on field-based data collection [118]. From a theoretical sampling standpoint, they also maximize diversity across several key project characteristics, while remaining comparable due to similar participation structures and levels of project maturity. As shown in Table 4.1, the projects discussed in the current work originated from diverse scientific domains in a variety of organizational structures. They operated at geographic scales ranging from local to global, with average contribution rates ranging from well under 7,000 data points per year to over 7,000 data points per hour.

The process of collecting and processing field data in a citizen science project is summarized in Figure 4.1. We describe a general workflow, noting that each citizen science project has its own specific variations to accommodate.

4.3.2 Task Execution

Definition. The monitoring task is usually defined by a project coordinator/domain expert, sometimes with participant input. Precise task design is critical for data quality and meaningful evaluation metrics [101]. Task definitions generally include step-by-step protocols for collecting data and explanations of targeted parameters.

For example, the CoCoRaHS task includes reading a rain gauge and (in winter months) melting snow samples to measure snow water equivalent. The River Watch task follows a more complex protocol based on professional monitoring processes, including readings with a chemical probe (sonde). Ideally, the task definition is digitized and stored in the project archive. In practice, it is documented primarily in the form of instructions to participants.

Assignment. Once volunteers agree to participate and have completed any necessary training, task assignment usually includes selecting the location(s) for collecting data. Participants typically make observations in areas near their home or school due to the practical constraint that volunteers must be able to readily access observation sites.

Participants may be assigned specific individual sites to monitor (River Watch); or they may report on data for established shared observation sites, such as permanent plots for plant species (Mountain Watch); or they may select a location in their yard (CoCoRaHS, The Great Sunflower Project); or they may make observations anywhere the phenomenon of interest is present (eBird and numerous others). In most cases, observation locations chosen by participants are resolved to a latitude and longitude based on a street address or a pin dropped on an online map.

The assignment step may also include training and/or distribution of equipment, where appropriate. In some projects, training is self-paced using online materials; in others, in-person workshops provide opportunities to learn more complex participation processes. For a few projects, training is limited to instructions for data entry: through self-selection, eBird participants are typically already “trained” in key skills for bird detection and identification, and need only learn the conventions of the eBird system.

Completion. Once protocols are defined and locations assigned, it is up to the volunteers to head out into the field to collect data. This results in time series of sampling events or observations. As noted in Chapter 1, an observation is is “the intersection of a person, a bird, a time, and a place" for eBird. Similarly, a River Watch sampling event is the combination of a sampling team, a time, and a predetermined site along a stream.

While CoCoRaHS and River Watch incorporate measurements from specialized sensors into their workflow, eBird, the Great Sunflower Project, and Mountain Watch each require nothing more than field observations written down on paper. Very simple task definitions are increasingly common among citizen science projects, because most organizers report a direct tradeoff between task complexity and volume of data contributions.

4.3.3 Data Management

Digitization. Ideally, all information for each observation would be uploaded into a centralized system immediately as a single report for instant validation and change tracking. In practice, the conversion of data from field notes into digital form often does not happen instantly. Data entry is sometimes substantially delayed because participants consider it an unpleasant or undesirable task; this is a universal challenge for projects that require a separate data entry step. In some cases, participants intentionally delay data submission due to concerns over data visibility for sensitive species and breeding animals.

It might seem that replacing paper-based data entry with a mobile app would streamline the process and facilitate instant validation and provenance tracking, but there are some notable barriers to consider. Some contributors will always be more comfortable manually recording data in the field, or hard copy record retention may be required for quality assurance or legal purposes, making mobile entry an unwanted extra step. Not all contributors own smartphones or other technologies such as GPS devices. In some projects, the primary contributor group is older adults, the demographic with lowest smartphone adoption [104]. Under a variety of circumstances, bulk upload can be the most feasible way to entice volunteers to contribute larger volumes of data.

Collecting data in the field is subject to the conditions of the field. Technical constraints can limit the usefulness of smartphones (e.g., off-grid usage must be accommodated), but sometimes using electronics is simply impractical due to screen glare, inclement weather, or incompatibility with the flow of activities. In addition, most projects are poorly positioned to manage an additional platform (or three) for mobile data entry. Although the use of HTML5 can facilitate cross-platform deployment [100], the mobile workflow adds complexities that can be challenging to address. Further, external partners may not be able to adapt to new technologies.

Even when new technologies can be incorporated, integrating the collected data into a single submitted record is still challenging. Unannotated digital photos may require later manual matching to field data. For measurement projects like River Watch, chemical sensor readings must be transcribed manually (though newer sensors will support direct data transfer). When lab samples are involved, the original field report must be associated with a lab report created after the fact.

Review. In order to ensure the quality of the project results, incoming data is often reviewed to the extent feasible. Large projects like eBird may use algorithmic flagging of outliers – as defined by the data themselves, where possible [45] – to automatically identify data points that require expert review. Review is conducted by domain experts; for example, eBird's network of approximately 650 volunteer reviewers use both online and offline tools and resources for data review. Most reviews are readily completed within the custom review interface, but during spring migration, when an especially high number of records are flagged each year due to early movement of some species, bulk review is more feasible. Reviewers sift through data in Excel to manage these large batches or identify more nuanced data problems, and then update records via eBird's online review interface (re-upload of edited records is not supported.)

In smaller projects without resources for such systems, data filtering algorithms are enacted by hand, often with Excel. These manual procedures are rarely adequately documented to make the data processing itself repeatable.

For example, River Watch data is reviewed by a core team of 3-4 staff. While the project's custom ICT does support some simple range validation checking, making corrections to bulk data while importing it has been a cumbersome task. Thus, most of the review to identify and remove outliers still happens in Excel, and reviewers generally prefer to wait to import anything until after they have fully reviewed the data they receive. As a result, potentially valuable information about the review process is lost, as is the ability to easily reverse a decision to discard an outlier when reversal is warranted.

Exchange. An important part of enabling data use is facilitating exchange with third parties. Data are often exchanged using a standard or agreed-upon format. For example, River Watch monitoring data is sent annually for incorporation in the Minnesota Pollution Control Agency's master database, using an Excel spreadsheet with a layout based on the STORET standard. In addition to less formally structured download files, eBird data packages are made available in Bird Monitoring Data Exchange (BMDE) format, an extension of Darwin Core metadata standards, which was collaboratively developed with the Avian Knowledge Network to permit data exchange among partners.

In some cases, third parties may comment on or even modify data within their systems as part of an ongoing review process. This is certainly true for River Watch, and it has traditionally been difficult to ensure that these changes are replicated across databases. This challenge is addressed in the new data model proposed in this chapter.

4.3.4 Task Refinement

Finally, as the project evolves, the core tasks will often be refined. A new measuring tool may become available, for example, or a task is simplified to make it accessible to a larger audience. Data users and contributors often request modifications to protocols to better fit their needs. For example, in 2012 River Watch (and other similar programs in Minnesota) switched from measuring water clarity with T-Tubes to using Secchi tubes. The protocol and data structure is nearly the same, but Secchi data is treated as a different parameter for provenance purposes. Although eBird's three primary protocols remain unchanged since the project launched, 20 additional protocols appear in the data set, most of which address specific needs for projects coordinated by partner organizations.

Coordinators may also find that the way volunteers execute tasks differs from their expectations, leading to a retooling of procedural details to support more consistent task completion. For example, The Great Sunflower Project switched from 30-minute sampling on Lemon Queen (*Helianthus annus*) sunflowers exclusively (2008), to 15-minute sampling (2009), to 15-minute sampling on Lemon Queens plus a selection of common garden species (2010), to 5+ minute sampling for any flowering garden species (2013). This is neither atypical nor slow development for a new participation protocol in field-based citizen science.

Supporting changing task definitions is a common challenge for ICT in citizen science. Without careful planning, project development can be stymied by static platforms that cannot adapt to changing project needs without additional (often unavailable) programming effort. Even more flexible data models typically fail to retain the information necessary to evaluate historical data properly. This can lead to erroneous assumptions about data and misinterpretations that can have serious consequences, e.g., inappropriate land use recommendations for managing endangered species habitat. Managing changes to task definitions is another challenge addressed by our proposed data model.

4.3.5 Key observations

Field monitoring data often starts its life “offline” and may also be reviewed and edited offline. Despite evolving ICT capabilities, it is not reasonable to expect that all contributors will use a custom system for all editing and review practices. Especially when third parties are involved, there is simply no way to ensure that every important revision to the data will happen inside of a project’s ICT system.

With this in mind, we suggest that rather than building a custom platform that supports every conceivable data review operation, it may be more valuable to build a system that is robust against *multiple imports and exports* to and from external formats. That is to say, data import is more than just a bootstrapping feature until developers can implement a full-featured data management platform, then train contributors to perform all data modification within it, where changes can be directly monitored. Instead, the ability to import new *and modified* data from external formats is, and will remain, a core part of the workflow. This can be due to contributor preferences, and to parts of the workflow not fully under project control, like data exchange with third parties.

In particular, the good old spreadsheet is how many scientific project contributors prefer to work with data [89]. Normalized data models and sophisticated apps are not necessarily seen as useful, even if there is a demonstrable overall benefit. Volunteers are rarely eager to learn new software, and data management tasks are a hurdle to participation for many individuals. Generally, our internal data model should account for the needs of contributors who are unconcerned about internal data models and just want to participate in science.

Certainly, there is great potential for mobile devices to be harnessed as a way to improve data quality and submission rates[2] . For River Watch, allowing on-the-spot data entry via a mobile application would eliminate a couple of data transferring steps and opportunities for data entry errors from the process. In addition, a mobile app could provide feedback as soon as data is collected - while there is still an opportunity to take another measurement.

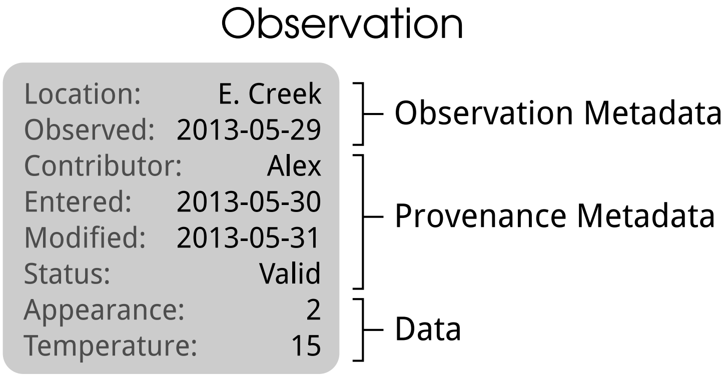
However, mobile technology is not a panacea, and in many projects, mobile entry should not be the only contribution option. For example, Kim et al. found that barriers to mobile adoption persist even when projects are provided with a free user-friendly campaign authoring tool [47]. Wiggins and He studied iNaturalist community logs and found that data submitted through the mobile app often had *lower* quality along some dimensions than data submitted through the website [120]. A more holistic approach should incorporate new technologies where appropriate (and possible), but also allow for more traditional ways of handling data, with the understanding that technology is just one part of the larger process [118]. Toward that end, we propose a new data model that flexibly integrates data from both bulk import and mobile field entry workflows.

4.4 Proposed Data Model

As we have discussed, data management for field-based observation is a nontrivial task. Most existing general-purpose platforms do not track changes to data, and wikis are not particularly well suited for field observation data. Even tailor-made platforms struggle with adequately maintaining history. An ideal data model would track changes to data and task definitions, allowing accurate analysis of historical data. Importantly, *the model must handle the data import task robustly and repeatedly*, by matching incoming records to data already in the database.

We propose a novel data model that has these characteristics, and hope that it will be useful to implementers of ICT for citizen science projects. We demonstrate the derivation of the new model by way of example. As a starting point, Figure 4.2 shows a simple “single-table” model for storing incoming observations.

[2] For example, the eBird team have noted that the BirdLog mobile app has dramatically increased submissions by core contributors.



Location:	E. Creek
Observed:	2013-05-29
Contributor:	Alex
Entered:	2013-05-30
Modified:	2013-05-31
Status:	Valid
Appearance:	2
Temperature:	15

Figure 4.2: Example *single-table* model for representing an observation. Note that "single-table" here refers only to the structure of the observation data, as there would typically be separate Contributor and Location entities, which are represented here as text values for simplicity.

The model contains a number of metadata attributes that are important for provenance, as well as the actual recorded values for the observation. Note that there are three different types of data being stored: Observation metadata, describing when and where the observation took place (Location, Observed); Provenance metadata, tracking the process of entering and maintaining the record within the ICT (Contributor, Entered, Modified, Status); and Data, or the actual values being reported per the task definition.

The single-table model conceptually matches an intuitive understanding of the task definition, and could easily be implemented in web and paper forms for entering the data. This database model is used by a number of projects, including CoCoRaHS, but can be inflexible to evolving project needs.

For example, a subset of River Watch participants now collect both precipitation and frost depth information during the winter months when streams are frozen [100]. The precipitation data is a natural fit for CoCoRaHS and so is shared with the program. However, there is currently no place in the CoCoRaHS database to record frost depth, which falls outside of the original scope of the project. Asking CoCoRaHS staff to add additional database columns and interface elements for the sake of one subcommunity is not reasonable, so as a workaround, these additional data were submitted to CoCoRaHS in the "comments" field, limiting its usefulness.

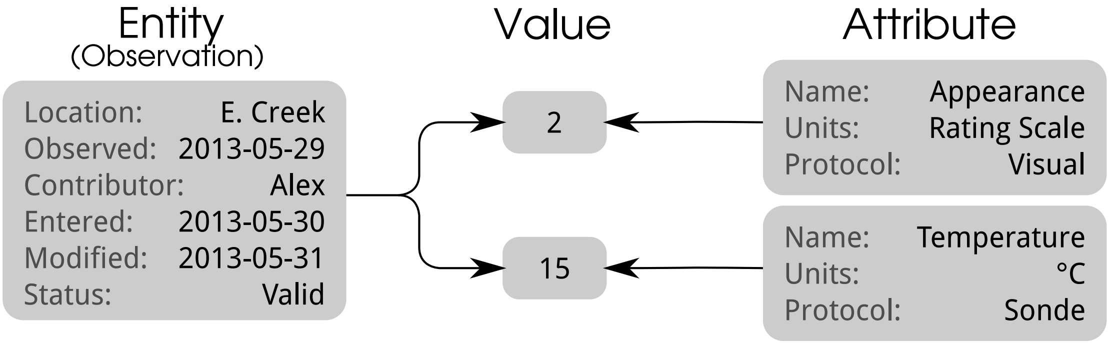


Figure 4.3: Simplified *Entity-Attribute-Value* model for field observations. Each rectangle represents a row in a table. Arrows represent parent-child relationships; some primary and foreign key columns are not shown.

In general, whenever a project task definition expands to include additional data fields, the project will need to have a developer add additional columns to the database and update hard-coded application logic. If an existing parameter definition changes in a way that will affect the meaning of the associated values, old data is either "upgraded" via computational means, or a subtle break in the data between the "old" and the "new" data is created.

4.4.1 EAV: Tracking Changes to the Task Definition

The entity-attribute-value model (shown in Figure 4.3) is a common way to address these flexibility issues. The EAV model is necessarily used by general-purpose platforms like Ushahidi to support custom task definitions. However, it is also useful for project-specific ICT systems.

For example, eBird's internal database model includes a *Checklist* (Entity), *Species name* (Attribute), and *Observation* (Value)[3] , the latter being the number of birds seen for a given species at the date and place represented by the checklist. Similarly, River Watch and other water quality projects' data are typically structured as a series of *Sampling Events* (Entities), each containing *Results* (Values) for a number of predetermined (but flexible) *Parameters* (Attributes).

By managing attribute definitions in their own table, systems designers can allow

[3] Note that "observation" as used elsewhere in this work refers to the entity (or checklist in eBird's case).

more flexibility to customize and maintain the task definition as needed. This table can also store useful metadata about each parameter (e.g., units and sampling methods) rather than hard-coding it in the application. When implementing EAV generally, it can sometimes be a challenge to determine which items to implement as attribute-values and which to leave as "normal" fields on the entity. In this case the distinction is relatively simple: metadata should generally be defined as part of the entity, while the observation or measurement data should be implemented as attribute-value pairs.

The EAV approach can also be applied as a simple but effective way to "version" task definitions, simply by creating additional attributes, though this capability is rarely exploited. New incoming observations could be associated with new attribute definitions, while leaving the existing data as-is. Where possible, it may be useful to define a mapping from the old values to the new values, but this can be done without actually changing prior data.

For example, when River Watch switched from T-Tubes to Secchi Tubes, one of the project coordinators was able to create a "Secchi Tube" parameter definition without developer intervention. While there was general agreement that the parameters should be kept separate in the database, River Watch organizers and participants wanted to evaluate their Secchi and T-Tube data on the same chart, as if they were the same parameter. This was facilitated by defining a relationship between the old and new parameters, indicating that they were numerically equivalent and could be graphed on the same scale.

4.4.2 ERAV: Tracking Changes to Observation Data

The EAV model is useful for supporting and tracking changes to the task definition. However, as we noted, individual observation records can also be changed during the review process, and these changes should ideally be tracked as well. With the model in Figure 4.3, if data is modified, the only indication that anything has changed is the *modified* column; the replaced data is lost entirely.

In a wiki or similar versioning system, a *version* field on the observation could be incremented whenever data change, automatically marking the previous version as *deleted* on each new save.[4] However, this assumes that revisions are done sequentially,

[4] Note that deleted is really just a value for the *status* field; an actual SQL DELETE would cause

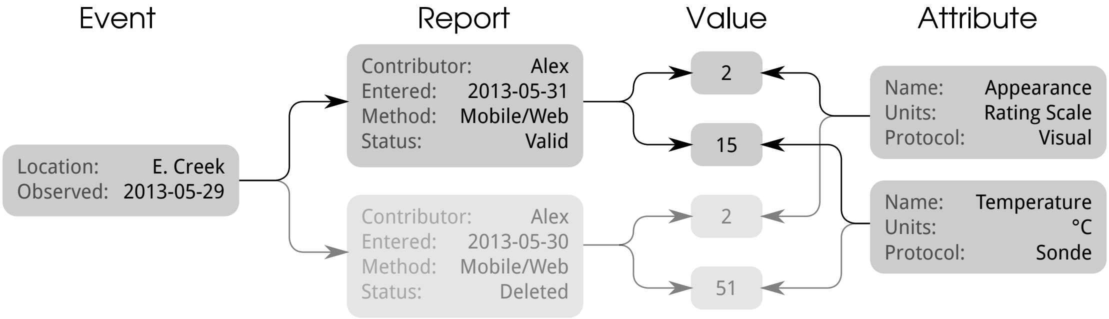


Figure 4.4: The Entity-Record-Attribute-Value Model

within the system, and that there is always only one active (non-deleted) version of the observation. As noted earlier, there can be two or more records created for the same observation, especially if there is lab work involved. We need a way to group these into a single entity for analysis, while maintaining the separate provenance information for each record.

To accomplish this, we propose a new model, ERAV, which separates the event entity (the actual observation) from its provenance record. The *entity* retains metadata about the observation (e.g., site, date observed), while the metadata about provenance (e.g., observer id, date/time entered, review status) is moved to a separate *record* table. *This approach enables maintaining multifaceted provenance data for the same monitoring event.*

As Figure 4.4 shows, whenever a record is changed within the system, a new version is created and the prior one is marked as deleted, much like in a typical wiki. Note that an explicit *last modified* column is no longer needed, as it can be derived from the *entered* date of the most recent record. As with EAV, the actual observations/measurements are recorded as attribute-value pairs. Importantly, these attribute-values are internally associated with each record, rather than the entity.

To the casual data user and for most data analysis purposes, the entity/record distinction is unimportant. For many common use cases, the attribute-values can be displayed as if they were directly associated with the entity. This trick allows provenance metadata for various attribute values to be maintained internally as part of each record, while presenting a single conceptual *event* entity for analysis purposes.

historical data to be lost, contrary to the goals of the versioning system.

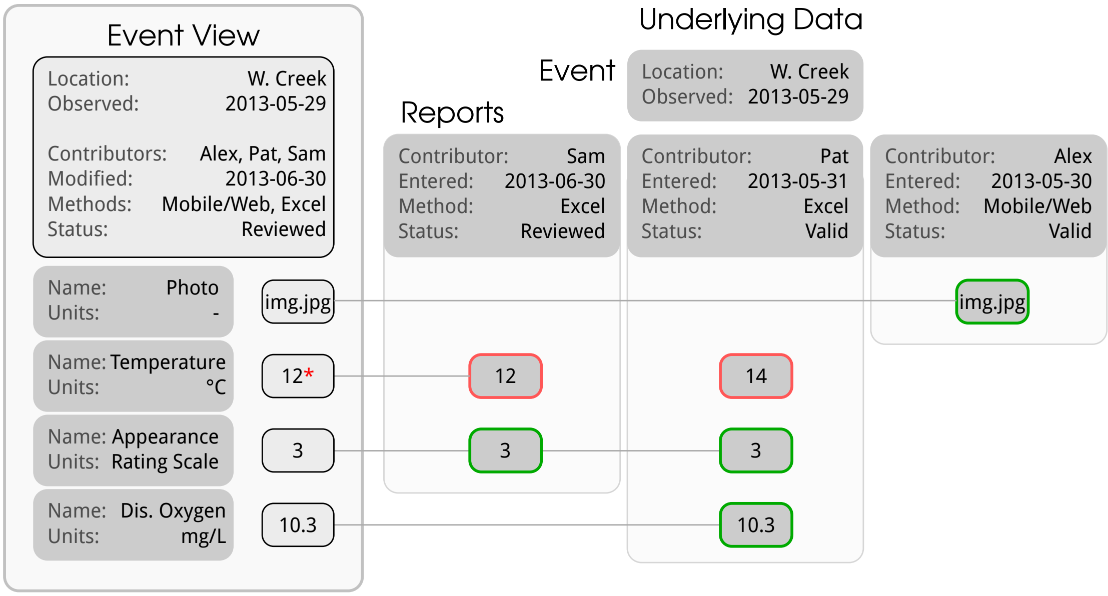


Figure 4.5: Resolving record information into a single "event" entity view. Some of the event data was imported from Excel more than once, leading to ambiguity. (For clarity, only currently active records are shown.)

4.4.3 Practical Concerns

"Merging" Records & Handling Conflicts

When information about an entity is created or modified outside of the system and then imported again, we cannot necessarily assume that any existing records for the entity should be deleted. For example, the incoming record may contain only supplemental data or a single additional attribute that was previously missed by mistake. Records can also contain lab data associated with corresponding field observations, which would previously have been manually merged, obfuscating provenance metadata in the process. In the new model, the contributor can simply upload the new file; if the observation metadata matches, new records are generated with the additional attribute-values, which are displayed together with the previously entered values as though all were directly associated with the entity. This feature can be thought of as "merging" records, though the actual data is not combined internally.

Even when the new data contains complete observation records, automatically deprecating older records may not be desirable. Certainly, when reviewers have already validated a bulk dataset offline, they prefer to avoid manually confirming each change, and instead want the system to re-import and accept their new values without question. On the other hand, if a third party conducted the review, the person responsible for re-importing the data may want to "re-review" each change separately. The main point here is that the appropriate time to resolve differences between versions is context, task, and contributor-dependent, so *the model must allow for ambiguities.*

This allowance can lead to discrepancies between attribute values for active records, particularly if the incoming data set is an update to existing entities incorporating data that was reviewed offline. The terminology of the conceptual model provides a convenient way to describe this situation: *conflicting reports.* This term can be used to alert contributors of the situation, where appropriate, but forcing contributors to deal with conflicts as soon as they are discovered is not always practical. Instead, the system may need to allow for ambiguities remaining in the dataset indefinitely. *This is the key difference between this model and traditional wiki models.*

Figure 4.5 demonstrates a number of possible outcomes when multiple imported records describe the same entity. If two active records for an entity contain values for different attributes, the data is merged without issue. Even if the records contain values for the same attributes, the data can be merged as long as the values are also the same. However, conflicting records must be handled differently according to the use case and skill level of each contributor.

For those only interested in using the data for analysis, conflicts can be smoothed over with a simple heuristic, as demonstrated by the Temperature value in Figure 4.5. By assigning a relative *authoritativeness* value to each record (e.g., the most recently *entered* record can often be treated as most authoritative), we can defer to the value from the most authoritative active record containing that attribute whenever there is a conflict. More complex workflows might assign priorities to different record status values (e.g. Provisional, Verified), allowing for more fine-tuned ordering and conflict resolution. This would also permit greater transparency for data use by third parties.

Identifying the Relevant Entity for an Incoming Record

So far, we have not dealt with the issue of matching incoming records to their associated entities. The ideal approach would be to include each entity's unique ID in exported files, and require that only appropriately annotated files be used for offline editing. Then, when a batch of records was (re-)imported, the embedded entity IDs could be used to match those records to the existing entities.

However, we cannot always control the format of batch files for upload, nor can we always ensure that updated files will contain the entities' unique identifiers used within our system. This is especially true for file formats controlled by third parties and standard exchange formats. Fortunately, there is usually a workable *natural key* for each entity we can use to match incoming records to entities. For example, assuming only one individual is monitoring a given site once a day, we can assume that any records for the same site and date should be associated with the same event entity[5] . The actual natural key used would project-specific, but the concept can be applied generally. It is important to be able to identify a usable natural key: if the key is too general, unrelated records will be inappropriately merged as if they were the same entity, but an overly specific key will prevent merging.

Obviously, this strategy works only as long as none of the fields in the natural key need to be updated. If an entity is entered with e.g. the wrong date, uploading a spreadsheet with the correct date would simply create another entity unassociated with the old data. One possible workaround might be to step back and compare the entire date range of the contributors' existing data against the date range of the uploaded file and point out any obvious discrepancies. Similarly, there is no explicit support for deletion: if a reviewer deletes a row or column from an exported spreadsheet, they might reasonably expect the same data to be deleted from the associated entities upon re-import. As Figure 4.5 shows, this action would be instead interpreted as a partial update by the model.

[5] The natural key in this case is analogous to common protocols for generating an "Activity ID" in professional water quality monitoring databases, but does not require a separate database field.

4.5 Reference Implementation

I have provided the source code for a generic implementation of ERAV as `vera`, which is available on the Python Package Index and on Github [6] . `vera` relies on and extends the wq framework, an open-source suite of Python and JavaScript libraries I built to support River Watch and other citizen science and crowdsourcing projects [100].

The core of `vera` is a collection of Django-powered database models. As Figure 4.6. shows, the `Event`, `Report`, `Parameter`, and `Result` tables directly correspond to the Entity, Record, Attribute, and Value concepts in ERAV. `vera` also provides three additional tables that aren't critical to the ERAV concept but are still useful:

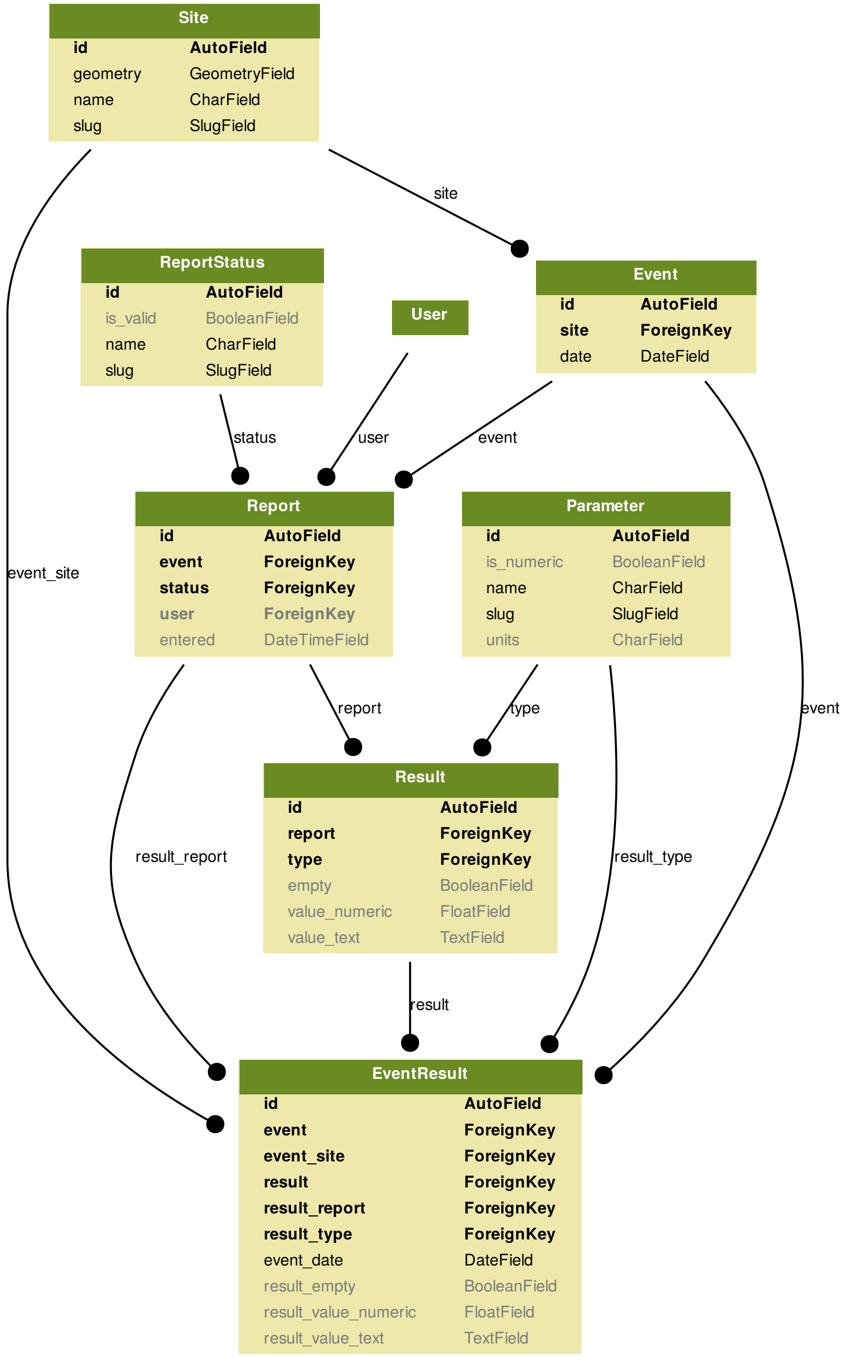
- The `Site` model allows for the definition and maintenance of pre-defined monitoring locations. River Watch and CoCoRaHS both make use of pre-assigned sites, but projects that allow ad-hoc observations may want to store location information on the `Event` model instead.
- The `ReportStatus` model allows for the customization of various valid and invalid record statuses, as well as their relative priority or "authoritativeness".
- Finally, the `EventResult` model provides a denormalized copy of all of the currently active `Result` data. Without this model, a query on the database must use extensive joining and sorting to ensure that the listed `Parameter`s for each `Event` are represented by the `Result` from the highest-priority valid `Report`. This query is not practical for interactive data exploration. `vera` automatically recomputes and updates the `EventResult` table whenever an `Event`, `Report`, or `Result` is modified. `vera` also provides out-of-the box charts and spreadsheet exports for the `EventResult` model via the Django REST Pandas API [7] .

The core concepts of ERAV are provided by the relationships between the tables in `vera`, but the tables themselves likely do not capture all of the metadata needed for most real-world projects. To facilitate use in a variety of domains, each Django model in `vera` is designed as an abstract base class with a default implementation that can be "swapped" for a custom definition without breaking the foreign key relationships[8] .

[6] `https://github.com/wq/vera`

[7] `https://github.com/wq/django-rest-pandas`

[8] `https://github.com/wq/django-swappable-models`



Site	
id	**AutoField**
geometry	GeometryField
name	CharField
slug	SlugField

ReportStatus	
id	**AutoField**
is_valid	BooleanField
name	CharField
slug	SlugField

User

Event	
id	**AutoField**
site	**ForeignKey**
date	DateField

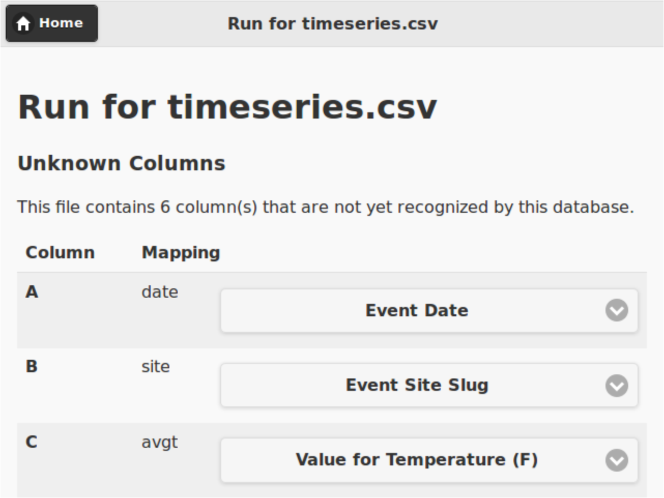
Report	
id	**AutoField**
event	**ForeignKey**
status	**ForeignKey**
user	**ForeignKey**
entered	DateTimeField

Parameter	
id	**AutoField**
is_numeric	BooleanField
name	CharField
slug	SlugField
units	CharField

Result	
id	**AutoField**
report	**ForeignKey**
type	**ForeignKey**
empty	BooleanField
value_numeric	FloatField
value_text	TextField

EventResult	
id	**AutoField**
event	**ForeignKey**
event_site	**ForeignKey**
result	**ForeignKey**
result_report	**ForeignKey**
result_type	**ForeignKey**
event_date	DateField
result_empty	BooleanField
result_value_numeric	FloatField
result_value_text	TextField

Figure 4.6: `vera` database schema with denormalized EventResult table



Run for timeseries.csv

Run for timeseries.csv

Unknown Columns

This file contains 6 column(s) that are not yet recognized by this database.

Column	Mapping	
A	date	Event Date
B	site	Event Site Slug
C	avgt	Value for Temperature (F)

Figure 4.7: Example Django Data Wizard Interface

`vera` would not be complete without robust support for data import from external formats. This is provided by the Django Data Wizard library [9] , which I recently refactored to work with or without `vera`. The unique feature of Django Data Wizard is that it allows novice users to map spreadsheet columns to "real" database fields *or* to ERAV and EAV attribute definitions without needing to know the details behind the scenes, as Figure 4.7 shows.

The "new" River Watch website [10] makes extensive use of all of the above technologies. The `vera` model allows coordinators from multiple independent OCS projects to define their own monitoring protocols. The model is also robust enough to support a two-way database synchronization with CoCoRaHS, allowing volunteers who are members of both projects to enter and explore all of their data in the River Watch site.

4.6 Discussion

As we have demonstrated, the proposed ERAV model facilitates useful data integration tasks that are difficult or impossible to accomplish in previous data models. The model is conceptually straightforward, with intuitive underlying concepts. ERAV builds off of

[9] `https://github.com/wq/django-data-wizard`

[10] `https://river.watch`

the existing EAV model to implement very flexible systems that can adapt to changing project needs while preserving data provenance.

Like EAV, ERAV is much more complex than a single-table approach. ERAV has the additional complexity of allowing multiple active versions of the same data, as well as historical versions, to be present in the database at the same time. This has negative implications for system performance[11] that can be mitigated through various technical means such as indexing, caching, and denormalization to a warehouse table for analysis. However, it may also make it more difficult for new developers to quickly understand and adapt an existing project's software. In particular, the multi-table database structure is somewhat less self-documenting, and effectively requires the application software to properly maintain it. Similarly, the software itself is dependent on the database to dictate the interface layout - a useful but occasionally befuddling feature familiar to designers of EAV systems.

Nevertheless, we argue that the flexibility and provenance capabilities ERAV provides are valuable enough to merit the additional complexity for many, if not most, OCS projects. More broadly, ERAV is likely to be useful in cases where:

1. Structured data is being exchanged and revised between multiple parties or data management platforms,
2. The selected (or de facto) exchange format does not include complete provenance information, and
3. The entities being described (i.e. events) can be uniquely identified with a stable natural key that does not need to be centrally assigned.

There is a need for more empirical evaluation to better understand the appropriate interfaces for representing "conflicting reports" to contributors of varying skill levels and expertise. We also note the potential of leveraging other common uses of Excel for review that are missed when only reading cell values for import. For example, River Watch reviewers often use various highlight colors to flag questionable values, but this information is lost after the spreadsheet is ingested. Finally, more technical work is needed to address cases like deletion that the existing model does not handle well.

[11] Concerns about performance were among the primary reasons CoCoRaHS initially chose not to utilize an EAV schema.

Recent advances in ICT can help improve tracking of revisions and retention of provenance metadata for citizen science projects. However, it is important to keep in mind that participants' expectations, skill levels, and preferred software tools may not match those idealized by computer scientists. For many projects, building complex data review tools within custom ICT would be over-engineering with minimal benefit. It is often more appropriate to enable bulk import processes, not only to facilitate ingestion of historical records, but as part of the ongoing development and improvement of the quality of the project. Retaining provenance metadata can be a challenge, and our proposed model is an important step toward enabling that capability.

Chapter 5

Conclusion

5.1 Summary of Contributions

This thesis provides a thorough examination of how data quality can be understood, measured, and improved in observational citizen science. The projects studied are reasonably representative samples of a wide range of OCS projects. Many of the lessons learned are generally applicable to crowdsourcing and peer production platforms, as well as to professional scientific data workflows. The key contributions of this thesis are as follows.

Understanding Data Quality. In Chapter 2 and [101], I demonstrated that data quality in OCS is inherently tied to the practices for maintaining it. This is especially true for smaller OCS projects like River Watch that do not have the scale necessary for automated quality control to be viable. But more fundamentally, some notions of quality are inherently uncomputable. This calls for careful consideration when attempting to computerize OCS workflows. Computer systems for OCS should be designed with the understanding that technology is just one part of a broader toolset of practices and perspectives.

Measuring Data Quality. In Chapter 3 and [102], I showed that volunteer retention is an effective way to evaluate data quality in OCS. OCS projects appear to attract a different volunteer population than online projects, perhaps because OCS has a more salient impact on day-to-day life. This may mean that existing practices for motivating and sustaining engagement in online peer production are not directly applicable to OCS.

Improving Data Quality. From my experiences with River Watch and CoCoRaHS, I derived a general description of the OCS workflow and validated it with additional insight from Andrea Wiggins. In Chapter 4 and [103], I describe the technical design and implementation of a data model that facilitates common "computer science" concerns (like provenance and revision tracking), while integrating with existing OCS workflows and remaining conscious of domain scientists' (relative) disinterest in complex technology. I also provide an implementation of the data model and note that it is used to facilitate exchange between River Watch and CoCoRaHS in particular.

5.2 Design Implications

The studies discussed in the previous chapters give rise to a number of general design implications for OCS and similar domains. We focus here on the HCI and CSCW implications, i.e. the design of technical systems for OCS. However, we note that successful OCS project design goes well beyond technical systems to include a variety of components from protocol definition to public outreach [65, 78].

Based on our analysis of River Watch and CoCoRaHS, together with a systemic review of OCS in general, we propose that OCS system implementations should prioritize the following features:

1. Track the full data quality process
2. Leverage intrinsic participant motivation
3. De-center technology

We elaborate on each of these points below. These implications can be viewed as layered constraints on the design of ICT systems for OCS data.

5.2.1 Track the full data quality process

Data quality in citizen science projects is a process, not only an attribute, and involves a number of components which are not necessarily present in the project dataset per se. Therefore it is important to understand quality concerns in context of the entire process when designing systems for maintaining and validating citizen science data. Ideally, systems should be built to support to track the entire data quality process in real time, rather than only serve as append-only repositories for the final dataset. Tracking the process will give insight into workflow challenges (as discussed in Chapter 2) as well activity patterns and long-term retention trends (as discussed in Chapter 3).

As discussed in Chapter 4, projects should track changes to project data as well as the protocols in place when the data was collected. Tracking these changes will help ensure accurate provenance metadata is available for each record, as well as provide a window into process improvement needs and individual skill development. This implies use of the ERAV model or a similar revision tracking system, subject to the remaining two constraints.

5.2.2 Leverage intrinsic participant motivation

Citizen science projects are often approached with the assumption that there is a trade-off between collecting scientifically useful data, and successfully motivating volunteers. However, this need not be the case. As Chapter 2 demonstrated, the twin goals of data collection and natural science education in River Watch are not in tension, and instead directly support each other. This is in part because the concern for quality data is itself a core part of the motivation teachers have for participating. Similarly, Chapter 3 demonstrated that many CoCoRaHS participants are intrinsically motivated from the start and remain active for a long time. This long-term retention is directly correlated with the definition of data quality in that domain.

Thus, technical systems for OCS should seek to leverage existing participant interest in the domain. While gamification may be useful for attracting new volunteer types (c.f. [7]), many OCS participants may not need any additional motivators beyond those inherent in the same data they are contributing. River Watch participants are naturally motivated to compare their stream's water quality against historical records, while CoCoRaHS participants love being able to look at the day's precipitation map to see how much rain their neighbors got. This interest and domain knowledge could be used to make data entry more robust - for example, by leveraging the ability for experienced River Watch participants to "just know" when a value is out of the expected range for a site.

With this in mind, a focus on data quality need not and should not preclude supporting more interpretive aspects of a citizen science project. Instead, directly addressing the motivations people have for participating in a project can increase interest and data quality as a consequence. In addition to providing export tools for expert users, significant value can be added by providing interpretive tools within the system so that volunteers can explore their data themselves. In addition, communication tools for sharing and discussing findings can increase engagement and provide a way to collect rich contextual information that can be used to improve the process. If volunteers were made aware whenever others use their data, they would likely be more motivated to contribute quality data.

Designing a system that explicitly supports interpretive goals will likely result in better data quality, because there will be more reason to interact with the website and review data on an ongoing basis. One way to do this is to give prominence to “soft” data like field notes and photographs in order to make the system more accessible and interesting, and to encourage teams to take good notes. Another approach is to assign meaningful URLs to each user-contributed artifact, making it easy to re-share them via existing social networks while preserving a central repository.

A successful system will engage participants without drawing attention away from the primary goals of the project, as discussed in the final constraint.

5.2.3 De-center Technology

Participants in OCS come with a wide range of interests and technology skill levels. In River Watch, most participants are fairly young, while support staff and coordinators are older. In CoCoRaHS, the dynamics are reversed, with many participants past retirement age and a (relatively) younger support network. But common to both projects (and many other OCS) is a strong intrinsic interest in the scientific domain, with an interest in technology playing a minor role if any.

By contrast, Wiki and OCS platforms are typically designed to center themselves as *the* project, sometimes with an explicit recommendation against use for any non-data entry tasks like social networking and community building. For OCS, however, data entry is a secondary task, separate from data collection, sampling, and other field-based tasks. It would not be too much of a stretch to say that while wikipedia.org and its subdomains *are* Wikipedia, cocorahs.org is better viewed as “just” the website for the CoCoRaHS project.

Thus, it is essential for designers of technical systems for OCS to not try to unnecessarily center their technology within a project. Certainly, this is a helpful mindset for any technology - as Norman himself would say, the ideal computer is invisible to its user [67]. However, this invisibility is particularly important within the realm of OCS.

5.3 Future Work

Moving forward, it will be important to extend this research in several ways. First, it would be useful to directly compare demographics, activity levels, and retention between CoCoRaHS and another large OCS project like eBird. Unlike CoCoRaHS, eBird does not have a fixed upper contribution limit, nor does it restrict volunteers to reporting from pre-assigned sites. We expect that these differences will have a measurable effect on the structure of contributions to the two systems.

While many of the constructs used in this research have clear analogues in peer production research, a more explicit comparison between the domains would facilitate broader generality. For example, a standard metric for early activity (i.e. first month contributions), as well as dropout (i.e. a year-long break in inactivity), would make it possible to directly compare retention outcomes between a variety of virtual and observational projects, as well as wikis and other platforms.

In addition, there needs to be more work evaluating the long tail of small, local scale observational citizen science projects like River Watch. These projects are notoriously hard to characterize from a computer science perspective, given the small `n` and the wide ranges of technologies used. However, they represent a large portion of the observational citizen science phenomenon, and are an important part of understanding its success.

The workflow model proposed in Chapter 4 could be used to frame subsequent research in this domain. In particular, we envision future work explicitly identifying which stages of the workflow are directly affected by proposed technological interventions. Similarly, we hope future projects will leverage and expand on the ERAV data model and its open-source implementation.

One key challenge I grappled with throughout the research that formed this thesis was this: How can I implement and validate technology solutions within existing OCS communities without unnecessarily disrupting their current workflow? Given the need to de-center technology in OCS, it is worth asking whether computer science *per se* is even the right epistomological foundation from which to study this phenomenon. In this regard, I am mindful of the questions posed by Baumer and Silberman ([4]), who suggest that the implication is sometimes not to design technology at all.

Given that observational citizen science is very often oriented around environmental and sustainability concerns, I would that future work in this domain might consider the field of *Sustainable HCI* as a starting point. Like other forms of HCI, sustainable HCI sits on the boundary between computer science and a multitude of related domains, and has thus grappled extensively with the tensions that future researchers of OCS may encounter.

5.4 Concluding Thoughts

Observational citizen science is a promising domain for expanding our knowledge of the natural world. As with other forms of computer-supported collaboration, data quality is an essential key to OCS project success. As this thesis demonstrates, data quality is also a powerful lens with which to better understand the intricacies of project work in this domain.

This research makes a significant first step toward filling the missing parts of our understanding of how OCS is different than other forms of open collaboration. Understanding the commonalities and differences will help more accurately determine which lessons in related domains apply to OCS and vice-versa. This knowledge contributes to a broader picture of how (and when) to support these communities with technology.

References

This thesis includes most of [101, 102, 103].

[1] Rodrigo Almeida, Barzan Mozafari, and Junghoo Cho. On the evolution of wikipedia. In *ICWSM*. Citeseer, 2007.

[2] Denise Anthony, Sean W Smith, and Tim Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of Wikipedia. *Hanover: Dartmouth College*, 2005.

[3] Avinoam Baruch, Andrew May, and Dapeng Yu. The motivations, enablers and barriers for voluntary participation in an online crowdsourcing platform. *Computers in Human Behavior*, 64:923–931, 2016.

[4] Eric PS Baumer and M Silberman. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2271–2274. ACM, 2011.

[5] Christopher Beirne and Xavier Lambin. Understanding the Determinants of Volunteer Retention Through Capture-Recapture Analysis: Answering Social Science Questions Using a Wildlife Ecology Toolkit. *Conservation Letters*, 6(6):391–401, 2013.

[6] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press, New Haven, CT, USA, 2006.

[7] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. Using gamification to inspire new citizen science

volunteers. In *Proceedings of the first international conference on gameful design, research, and applications*, pages 18–25. ACM, 2013.

[8] Dana Burr Bradley. A reason to rise each morning: The meaning of volunteering in the lives of older adults. *Generations*, 23(4):45, 1999.

[9] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. Why would anybody do this?: Understanding Older Adults' Motivations and Challenges in Crowd Work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2246–2257. ACM, 2016.

[10] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1101–1110. ACM, 2008.

[11] S van Buuren and CGM Oudshoorn. Multivariate imputation by chained equations: MICE V1. 0 user's manual. Technical report, TNO, 2000.

[12] Susan M Chambré. Volunteerism by elders: Past trends and future prospects. *The Gerontologist*, 33(2):221–229, 1993.

[13] Alice Ming Lin Chong, Tina Louisa Rochelle, and Susu Liu. Volunteerism and positive aging in Hong Kong: A cultural perspective. *The International Journal of Aging and Human Development*, 77(3):211–231, 2013.

[14] Robert Cifelli, Nolan Doesken, Patrick Kennedy, Lawrence D Carey, Steven A Rutledge, Chad Gimmestad, and Tracy Depue. The community collaborative rain, hail, and snow network: Informal education for scientists and citizens. *Bulletin of the American Meteorological Society*, 86(8):1069–1077, 2005.

[15] Ram A Cnaan and Toni A Cascio. Performance and commitment: Issues in management of volunteers in human service organizations. *Journal of Social Service Research*, 24(3-4):1–37, 1998.

[16] Cathy Conrad and Krista Hilchey. A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental Monitoring and Assessment*, pages 1–19, 2010.

[17] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *CHI '06*, pages 1037–1046, New York, NY, USA, 2006. ACM.

[18] Joe Cox, Eun Young Oh, Brooke Simmons, Gary Graham, Anita Greenhill, Chris Lintott, Karen Masters, et al. Doing Good Online: An Investigation into the Characteristics and Motivations of Digital Volunteers. In *12th International Conference of the International Society for Third Sector Research*, 2016.

[19] Martin Dittus, Giovanni Quattrone, and Licia Capra. Analysing volunteer engagement in humanitarian mapping: building contributor communities at large scale. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 108–118. ACM, 2016.

[20] Kimberly L Elmore, ZL Flamig, V Lakshmanan, BT Kaney, V Farmer, Heather D Reeves, and Lans P Rothfusz. mPING: Crowd-sourcing weather reports for research. *Bulletin of the American Meteorological Society*, 95(9):1335–1342, 2014.

[21] Environmental Protection Agency. *National Directory of Volunteer Monitoring Programs*, 2011. `http://yosemite.epa.gov/water/volmon.nsf`.

[22] Cameron Davidson-Pilon et al. Camdavidsonpilon/lifelines, June 2017.

[23] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L Cox. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2985–2994. ACM, 2014.

[24] Federal Geographic Data Committee and others. FGDC-STD-001-1998. *Content standard for digital geospatial metadata*, 1998.

[25] Eric H Fegraus, Sandy Andelman, Matthew B Jones, and Mark Schildhauer. Maximizing the value of ecological data with structured metadata: An introduction to

ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3):158–168, 2005.

[26] Karen Firehock and Jay West. A brief history of volunteer biological water monitoring using macroinvertebrates. *Journal of the North American Benthological Society*, 14(1):197–202, 1995.

[27] Paul Fugelstad, Patrick Dwyer, Jennifer Filson Moses, John Kim, Cleila Anna Mannino, Loren Terveen, and Mark Snyder. What makes users rate (share, tag, edit...)?: predicting patterns of participation in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 969–978. ACM, 2012.

[28] Yolanda Gil, Simon Miles, Khalid Belhajjame, Henela Deus, Daniel Garijo, Graham Klyne, Paolo Missier, Stian Soiland-Reyes, and Stephan Zednik. *A Primer for the PROV Provenance Model*. W3C, 2012. `http://www.w3.org/TR/prov-primer/`.

[29] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[30] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group*, 2010.

[31] Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[32] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[33] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. The rise and decline of an open collaboration system: How Wikipedias reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688, 2013.

[34] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, page 15. ACM, 2009.

[35] Carl Hartung, Adam Lerer, Yaw Anokwa, Clint Tseng, Waylon Brunette, and Gaetano Borriello. Open Data Kit: Tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 18. ACM, 2010.

[36] Brent J Hecht and Darren Gergle. On the localness of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 229–232. ACM, 2010.

[37] Wesley M Hochachka, Daniel Fink, Rebecca A Hutchinson, Daniel Sheldon, Weng-Keen Wong, and Steve Kelling. Data-intensive science applied to broad-scale citizen science. *Trends in ecology & evolution*, 27(2):130–137, 2012.

[38] Catherine Hoffman, Caren B Cooper, Eric B Kennedy, Mahmud Farooque, and Darlene Cavalier. SciStarter 2.0: A Digital Platform to Foster and Study Sustained Engagement in Citizen Science. In *Analyzing the Role of Citizen Science in Modern Research*, pages 50–61. IGI Global, 2017.

[39] J. Howe. The Rise of Crowdsourcing. *Wired Magazine*, 14(6):1–4, 2006.

[40] Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, and Licia Capra. Mapping community engagement with urban crowd-sourcing. In *Proceedings of When the City Meets the Citizen Workshop, Dublin, Ireland*, volume 4, 2012.

[41] Desislava Hristova, Giovanni Quattrone, Afra J Mashhadi, and Licia Capra. The life of the party: Impact of social mapping in openstreetmap. In *ICWSM*, 2013.

[42] Caroline Jay, Robert Dunne, David Gelsthorpe, and Markel Vigo. To Sign Up, or not to Sign Up?: Maximizing Citizen Science Contribution Rates through Optional Registration. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1827–1832. ACM, 2016.

[43] Joseph M Juran. *Quality control handbook*. McGraw-Hill, 1962.

[44] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. Early Activity Diversity: Assessing Newcomer Retention from First-Session Activity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 595–608. ACM, 2016.

[45] Steve Kelling, Jun Yu, Jeff Gerbracht, and Weng-Keen Wong. Emergent filters: Automated data verification in a large-scale citizen science project. In *Proceedings of Workshops at the Seventh International Conference on eScience*, pages 20–27. IEEE, 2011.

[46] Sunyoung Kim, Jennifer Mankoff, and Eric Paulos. Sensr: Evaluating a flexible framework for authoring mobile data-collection tools for citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1453–1462. ACM, 2013.

[47] Sunyoung Kim, Jennifer Mankoff, and Eric Paulos. Exploring barriers to the adoption of mobile technologies for volunteer data collection campaigns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3117–3126. ACM, 2015.

[48] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007.

[49] Masatomo Kobayashi, Shoma Arita, Toshinari Itoko, Shin Saito, and Hironobu Takagi. Motivating multi-generational crowd workers in social-purpose work. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1813–1824. ACM, 2015.

[50] Kathrin Komp, Theo Van Tilburg, and Marjolein Broese Van Groenou. Age, retirement, and health as factors in volunteering in later life. *Nonprofit and Voluntary Sector Quarterly*, 41(2):280–299, 2012.

[51] S Lakshminarayanan. Using citizens to do science versus citizens as scientists. *Ecology and Society*, 12, 2007. Response 2. `http://www.ecologyandsociety.org/vol12/iss2/resp2/`.

[52] Anne M Land-Zandstra, Jeroen LA Devilee, Frans Snik, Franka Buurmeijer, and Jos M van den Broek. Citizen science on a smartphone: Participants motivations and learning. *Public Understanding of Science*, 25(1):45–60, 2016.

[53] Jeffrey Laut, Francesco Cappa, Oded Nov, and Maurizio Porfiri. Increasing citizen science contribution using a virtual peer. *Journal of the Association for Information Science and Technology*, 68(3):583–593, 2017.

[54] Anna Lawrence. 'no personal motive?' volunteers, biodiversity, and the false dichotomies of participation. *Ethics, Place and Environment*, 9:279–298(20), October 2006.

[55] Jun Liu and Sudha Ram. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):11, 2011.

[56] Wendy Liu and Derek Ruths. What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, page 01, 2013.

[57] Roman Lukyanenko, Jeffrey Parsons, and Yolanda Wiersma. Citizen science 2.0: Data management principles to harness the power of the crowd. In *Service-Oriented Perspectives in Design Science Research*, pages 465–473. Springer, 2011.

[58] Kurt Luther, Scott Counts, Kristin B. Stecher, Aaron Hoff, and Paul Johns. Pathfinder: an online collaboration environment for citizen scientists. In *CHI '09*, pages 239–248, New York, NY, USA, 2009. ACM.

[59] Ding Ma, Mats Sandberg, and Bin Jiang. Characterizing the heterogeneity of the OpenStreetMap data and community. *ISPRS International Journal of Geo-Information*, 4(2):535–550, 2015.

[60] Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[61] Adina M Merenlender, Alycia W Crall, Sabrina Drill, Michelle Prysby, and Heidi Ballard. Evaluating environmental education, citizen science, and stewardship through naturalist programs. *Conservation Biology*, 30(6):1255–1265, 2016.

[62] Hannah J Miller, Shuo Chang, and Loren G Terveen. I LOVE THIS SITE! vs. It's a little girly: Perceptions of and Initial User Experience with Pinterest. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1728–1740. ACM, 2015.

[63] Julia Nerbonne and Kristen Nelson. Volunteer macroinvertebrate monitoring: Tensions among group goals, data quality, and outcomes. *Environmental Management*, 42:470–479, 2008.

[64] Greg Newman, Jim Graham, Alycia Crall, and Melinda Laituri. The art and science of multi-scale citizen science support. *Ecological Informatics*, 6(3):217–227, 2011.

[65] Greg Newman, Andrea Wiggins, Alycia Crall, Eric Graham, Sarah Newman, and Kevin Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012.

[66] E. Nicholson, J. Ryan, and D. Hodgkins. Community data - where does the value lie? assessing confidence limits of community collected water quality data. *Water Science and Technology*, 45:193–200, 2002.

[67] Donald A Norman. *The invisible computer: why good products can fail, the personal computer is so complex, and information appliances are the solution.* MIT press, 1998.

[68] Oded Nov, Ofer Arazy, and David Anderson. Technology-Mediated Citizen Science Participation: A Motivational Model. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[69] Ory Okolloh. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1):65–70, 2009.

[70] Morris A Okun and Amy Schultz. Age and motives for volunteering: testing hypotheses derived from socioemotional selectivity theory. *Psychology and Aging*, 18(2):231–239, 2003.

[71] Allen M Omoto, Mark Snyder, and Steven C Martino. Volunteerism and the life course: Investigating age-related agendas for action. *Basic and applied social psychology*, 22(3):181–197, 2000.

[72] Fabrizio Orlandi and Alexandre Passant. Modelling provenance of DBpedia resources using Wikipedia contributions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):149–164, 2011.

[73] Felipe Ortega and Daniel Izquierdo-Cortazar. Survival analysis in open development projects. In *ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development*, pages 7–12. IEEE, 2009.

[74] Leysia Palen, Robert Soden, T Jennings Anderson, and Mario Barrenechea. Success & scale in a data-producing organization: the socio-technical evolution of openstreetmap in response to humanitarian events. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 4113–4122. ACM, 2015.

[75] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 51–60. ACM, 2009.

[76] Jone L Pearce. *Volunteers: The organizational behavior of unpaid workers*. Routledge, 1993.

[77] Lesandro Ponciano, Francisco Brasileiro, Robert Simpson, and Arfon Smith. Volunteers' Engagement in Human Computation for Astronomy Projects. *Computing in Science & Engineering*, 16(6):52–59, 2014.

[78] Nathan R Prestopnik and Kevin Crowston. Citizen science system assemblages: understanding the technologies that support crowdsourced science. *Proceedings of the 2012 iConference*, pages 168–176, 2012.

[79] Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268. ACM, 2007.

[80] Reid Priedhorsky and Loren Terveen. Wiki grows up: Arbitrary data models, access control, and beyond. In *Proceedings of the Seventh International Symposium on Wikis and Open Collaboration*, page 6371. ACM, 2011.

[81] PRISM Climate Group, Oregon State University, 2017. created 11 Mar 2017.

[82] Jordan Raddick, CJ Lintott, K Schawinski, D Thomas, RC Nichol, D Andreescu, S Bamford, KR Land, P Murray, A Slosar, et al. Galaxy Zoo: An experiment in public science participation. In *Bulletin of the American Astronomical Society*, volume 39, page 892, 2007.

[83] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. Galaxy Zoo: Motivations of citizen scientists. *Astronomy Education Review*, 12(1), 2013.

[84] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925*, 2009.

[85] Sudha Ram and Jun Liu. Understanding the semantics of data provenance to support active conceptual modeling. In *Active conceptual modeling of learning*, pages 17–29. Springer, 2007.

[86] Red Lake Watershed District and Red River Basin Monitoring Advisory Committee. *Standard Operating Procedures for Water Quality Monitoring in the Red*

River Watershed, 8th edition, March 2011. http://www.redlakewatershed.org/waterquality/RLWD%20SOP%20Revision%208.pdf.

[87] Jason Reed, M Jordan Raddick, Andrea Lardner, and Karen Carney. An exploratory factor analysis of motivations for participating in Zooniverse, a collection of virtual citizen science projects. In *46th Hawaii International Conference on System Sciences*, pages 610–619. IEEE, 2013.

[88] Henry W Reges, Nolan Doesken, Julian Turner, Noah Newman, Antony Bergantino, and Zach Schwalbe. COCORAHS: The evolution and accomplishments of a volunteer rain gauge network. *Bulletin of the American Meteorological Society*, 97(10):1831–1846, 2016.

[89] David Ribes and Thomas A Finholt. Representing community: Knowing users in the face of changing constituencies. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 107–116. ACM, 2008.

[90] Dirk Riehle. How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *WikiSym '06*, pages 3–8, New York, NY, USA, 2006. ACM.

[91] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM, 2010.

[92] Mary Roth and Wang-Chiew Tan. Data integration and data exchange: Its really about time. In *Proceedings of the 6th Biennial Conference on Innovative Data Systems Research*. CIDR, 2013.

[93] Dana Rotman, Jen Hammock, Jenny Preece, Derek Hansen, Carol Boston, Anne Bowser, and Yurong He. Motivations affecting initial and long-term participation in citizen science projects in three countries. *iConference 2014 Proceedings*, 2014.

[94] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. Dynamic changes in motivation in

collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 217–226. ACM, 2012.

[95] Robert L Ryan, Rachel Kaplan, and Robert E Grese. Predicting volunteer commitment in environmental stewardship programmes. *Journal of Environmental Planning and Management*, 44(5):629–648, 2001.

[96] Henry Sauermann and Chiara Franzoni. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 112(3):679–684, 2015.

[97] Beth Savan, Alexis J. Morgan, and Christopher Gore. Volunteer environmental monitoring and the role of the universities: The case of citizens' environment watch. *Environmental Management*, 31:0561–0568, 2003.

[98] Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, 2009.

[99] Avi Segal, Ya'akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. Intervention Strategies for Increasing Engagement in Crowdsourcing: Platform, Predictions, and Experiments. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3861–3867. AAAI Press, 2016.

[100] S Andrew Sheppard. wq: A modular framework for collecting, storing, and utilizing experiential VGI. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 62–69. ACM, 2012.

[101] S Andrew Sheppard and Loren Terveen. Quality is a verb: the operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 29–38. ACM, 2011.

[102] S Andrew Sheppard, Julian Turner, Jacob Thebault-Spieker, Haiyi Zhu, and Loren Terveen. Never too old, cold, or dry to watch the sky: A survival analysis of citizen

science volunteerism. *Proceedings of the ACM on Human-Computer Interaction*, 1(2):94:1–94:21, 2017.

[103] S Andrew Sheppard, Andrea Wiggins, and Loren Terveen. Capturing quality: retaining provenance for curated volunteer monitoring data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1234–1245. ACM, 2014.

[104] A. Smith. Smartphone adoption and usage. Technical report, Pew Internet & American Life Project, Washington, DC, 2011.

[105] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. Data quality in context. *Commun. ACM*, 40:103–110, May 1997.

[106] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.

[107] Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014.

[108] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[109] Viswanath Venkatesh, James Y. L. Thong, and Xin Xu. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Q.*, 36(1):157–178, March 2012.

[110] Jakob Voss. Measuring wikipedia. 2005.

[111] Denny Vrandečić, Varun Ratnakar, Markus Krötzsch, and Yolanda Gil. Shortipedia: Aggregating and curating Semantic Web data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):334–338, 2011.

[112] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[113] Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12:5–33, March 1996.

[114] R.Y. Wang, V.C. Storey, and C.P. Firth. A framework for analysis of data quality research. *Knowledge and Data Engineering, IEEE Transactions on*, 7(4):623 –640, August 1995.

[115] Zhimin Wang, Hui Dong, Maureen Kelly, James A Macklin, Paul J Morris, and Robert A Morris. Filtered-Push: A Map-Reduce platform for collaborative taxonomic data management. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, volume 3, pages 731–735. IEEE, 2009.

[116] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One*, 7(1):e29715, 2012.

[117] A. Wiggins, R. Bonney, E. Graham, S. Henderson, S. Kelling, R. Littauer, G. LeBuhn, K. Lotts, W. Michener, G. Newman, E. Russell, R. Stevenson, and J. Weltzin. *Data Management Guide for Public Participation in Scientific Research.* DataONE, 2013.

[118] Andrea Wiggins. Free as in puppies: compensating for ICT constraints in citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1469–1480. ACM, 2013.

[119] Andrea Wiggins and Kevin Crowston. From conservation to crowdsourcing: A typology of citizen science. In *44th Hawaii international conference on System Sciences*, pages 1–10. IEEE, 2011.

[120] Andrea Wiggins and Yurong He. Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1548–1559. ACM, 2016.

[121] Andrea Wiggins, Greg Newman, Robert D Stevenson, and Kevin Crowston. Mechanisms for data quality and validation in citizen science. In *Proceedings of Workshops at the Seventh International Conference on eScience*, pages 14–19. IEEE, 2011.

[122] Wikipedia contributors. Wikipedia:ignore all rules. *Wikipedia, the Free Encyclopedia*, 2011. `http://en.wikipedia.org/wiki/Wikipedia:Ignore_all_rules`.

[123] C. C. Wilderman. Models of community science: Design lessons from the field. In *Citizen Science Toolkit Conference*, Cornell Laboratory of Ornithology, Ithaca, NY, 2007.

[124] John Wilson. Volunteerism research: A review essay. *Nonprofit and Voluntary Sector Quarterly*, 41(2):176–212, 2012.

[125] Ilene Wolcott, Dean Ingwersen, Michael A Weston, Chris Tzaros, et al. Sustainability of a long-term volunteer-based bird monitoring program: recruitment, retention and attrition. *Australian Journal on Volunteering*, 13(1), 2008.

[126] Chris Wood, Brian Sullivan, Marshall Iliff, Daniel Fink, and Steve Kelling. eBird: engaging birders in science and conservation. *PLoS Biol*, 9(12):e1001220, 2011.

[127] Taha Yasseri, Giovanni Quattrone, and Afra Mashhadi. Temporal analysis of activity patterns of editors in collaborative mapping project of OpenStreetMap. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 13. ACM, 2013.

[128] Dell Zhang, Karl Prior, and Mark Levene. How long do Wikipedia editors keep active? In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 4. ACM, 2012.

[129] Haiyi Zhu, Robert E Kraut, and Aniket Kittur. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–290. ACM, 2014.