Quality is a Verb: The operationalization of data quality in a citizen science community

S. Andrew Sheppard, Loren Terveen
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, Minnesota
sheppard@umn.edu, terveen@cs.umn.edu

ABSTRACT

Citizen science is becoming more valuable as a potential source of environmental data. Involving citizens in data collection has the added educational benefits of increased scientific awareness and local ownership of environmental concerns. However, a common concern among domain experts is the presumed lower quality of data submitted by volunteers. In this paper, we explore data quality assurance practices in River Watch, a community-based monitoring program in the Red River basin. We investigate how the participants in River Watch understand and prioritize data quality concerns. We found that data quality in River Watch is primarily maintained through universal adherence to standard operating procedures, but there remain areas where technological intervention may help. We also found that rigorous data quality assurance practices appear to enhance rather than hinder the educational goals of the program. We draw implications for the design of quality assurance mechanisms for River Watch and other citizen science projects.

Categories and Subject Descriptors

K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*
; K.6.4 [**System Management**]: Quality assurance

General Terms

Design, Human Factors, Verification

Keywords

citizen science, community-based monitoring, data quality

1. INTRODUCTION

Open collaboration is emerging as an effective way to generate valuable repositories of information by spreading the effort out across communities of volunteers [1, 4]. However, domain experts may question the quality of the information, since it is maintained by volunteers with unknown and varying levels of knowledge and skill.

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 [2]. 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 [16, 24].

Citizen science projects such as eBird [20], Citizens Environment Watch [17], and Pathfinder [11] can be considered forms of open collaboration. The quality of the data collected by volunteers in citizen science programs is crucial [13], especially when it is used by resource managers to make assessments of the environment and resulting policy decisions. Involving citizens in science provides the additional benefit of increased scientific awareness. In this case, citizen science serves a dual role of providing a relatively inexpensive source of data while meeting educational goals.

"Volunteer" need not mean "untrained", and the level of rigor required in a citizen science community depends on its emphasis. More interpretive approaches may accept that data will have lower quality, emphasizing instead awareness, participation, and educational aspects [17]. However, other 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.

In this paper, we investigate data quality concerns in the context of River Watch, a citizen science 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 open collaboration 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?

WikiSym'11, October 3–5, 2011, Mountain View, California

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 paper, we will first discuss related work and the background of River Watch, then present results from the interviews and implications for the design of online systems to promote data quality in citizen science communities like River Watch.

2. RELATED WORK

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 [21]. While data quality may seem to be an intuitive concept (e.g. "correctness"), it is impossible to propose a precise, universally meaningful definition [19, 21, 22]. 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.

Wand and Wang identify a number of common dimensions used in data quality research [21]. 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 [18]. Wang and Strong investigated quality from this external perspective, by conducting surveys of consumers of data [23]. In this paper, 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 Quality in Open Collaboration Systems

Open collaboration provides a unique domain for studying data quality concerns. These can be investigated from the external perspective, as in Giles' well-known study [6] that investigated the quality of articles in Wikipedia by having independent reviewers evaluate them. However, because in open collaboration the line between producer and consumer is blurred [19], it is also valuable to study data quality from the perspective of the community itself.

Stvilia et al. [19] 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 this work.

Our approach is more similar to that of Riehle [16], 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. [7]. 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.

2.3 Quality in Citizen Science

Citizen science provides a way for citizens to actively participate in decisions that affect them by collecting the data used to make those decisions. It is thus a form of open collaboration, especially when citizen volunteers are treated not only as collectors of data but as fellow scientists [9] and participate in defining scientific goals as well as executing them [11]. Conrad and Hilchey [3] 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, though they concur with Lawrence [10] that the various approaches to citizen science are not necessarily mutually exclusive categories.

A particularly large subset of citizen science projects fall under the umbrella of *community-based monitoring* [3]. For example eBird [20], Creek Watch [8], and Citizen's Environment Watch [17] involve volunteers in environmental monitoring activities, which fill in the gaps in professional or governmental datasets. Environmental data is valuable to government agencies in determining potential environmental hazards, so it is important that the data be of good quality.

The quality of data in community-based monitoring 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 [13]. Nerbonne and Nelson [12] 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 the latter, primarily internal approach to be more valuable in the context of River Watch.

Citizen science has long been recognized for its educational benefits [13, 17], 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.

[1] In this paper, "data quality" will be used to refer to both data and information quality.

	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 1: Roles in River Watch

Studies of citizen science often involve the creation of new applications, and the primary questions in such projects are often how to effectively engage volunteers. Kim et al. [8] however note the importance of obtaining data that is actually useful for scientific analysis when designing a new project. In this paper, we are studying an established citizen science community that collects data that is already being used by agencies to make decisions. We therefore focus on ways that technology can improve existing quality assurance processes.

3. BACKGROUND AND METHODOLOGY

3.1 Background

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 paper. However, they can be broadly categorized in the following different types:

- 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 paper.

3.2 Methodology

The first author has been involved with the River Watch project for several years, primarily by developing online tools to manage River Watch data. 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 [15].

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

[2] “River Watch” 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 [5]. However, for the purposes of this paper, 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://riverwatch.umn.edu`

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

there are multiple occasions for interaction with other participants. The first author 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.
- 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.

We also interviewed 15 participants from a number of distinct roles (see Table 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.

4. RESULTS

4.1 How is quality defined?

RQ1. How is quality defined in an open collaboration 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 [21]. 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. [21]. 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 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.

4.1.1 Reliability

Reliability refers to the trustworthiness of the data, and its usefulness for answering pertinent questions about the watershed. 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:

> "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.

4.1.2 *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 [21].

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. [13] 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" [22]. 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.

4.1.3 *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 [14] 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" [14].

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.

4.1.4 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.

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.

4.2.1 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 a 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.

4.2.2 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.

4.2.3 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".

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 current 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".

4.2.4 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 field-based citizen science projects from other open collaboration projects like Wikipedia, where article quality can theoretically be improved at any point in time.

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.

4.3.1 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.

4.3.2 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.

4.3.3 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.

5. DISCUSSION

While providing valuable insight into the quality assurance practices in River Watch, the results of this study also lead to a number of general lessons for citizen science projects. These lessons are discussed below, together with implications for the design of technological systems for promoting data quality in such programs and in River Watch in particular.

5.1 Lesson 1: Design for quality as a verb

Data quality in citizen science projects is a process, not only an attribute, and involves a number of components which are not all present in the data itself. Therefore it is important to understand quality concerns in context of the entire process when designing systems for maintaining and validating citizen science data.

From RQ1 we learned that SOPs are the primary way quality is maintained in River Watch. Thus, the website could be extended to provide in-depth information about SOPs in the form of training videos, photos, and interactive content. These would ideally include explanations as to why SOPs are the way they are (i.e. what errors are they trying to prevent), in order to promote understanding and not just memorization. While such content additions would be helpful, a more fundamental design shift is needed.

5.1.1 Implication: Track process, not just data

Computer systems for managing citizen science data can be most effective when they are built to support the entire process rather than only as repositories for the data. The current River Watch website acts primarily as a data management tool – a task it seemingly performs well enough. However, little information is recorded about the process. As discussed previously, data values which participants and coordinators don’t trust are simply thrown out, usually before they even make it to the database.

Conversely, in open collaboration systems like Wikipedia, much of this process information is tightly linked to the data itself, providing a valuable way to quantitatively evaluate the process[19]. Similarly, we believe that using a wiki-like approach to track the process of recording and maintaining citizen science data will prove a useful way to capture valuable domain knowledge and identify potential areas for improvement.

Utilize wiki motifs. The system can track corrections to the data while preserving the old values, similar to the “Revision history” in Wikipedia. Change notes provide an opportunity for coordinators to explain their justification for removing incorrect values, providing valuable domain knowledge that can be used to improve the ability of the system to automatically catch errors.

Discussions about questionable data happen largely offline. It would be valuable to provide an opportunity for teams, coordinators and users to discuss questions about specific data points and record how consensus was reached as to whether the data is likely to be accurate. This functionality would be analogous to a “Talk Page” in Wikipedia.

Compute existing QA metrics. Field duplicates are marked as such in the current system, but there is no domain-aware use made of the information. The website should automatically compute the Relative Percent Difference, and grossly inconsistent duplicates should be reported back to the team or even to the coordinators.

Similarly, calibration data should be uploaded together with the sampling data and strongly linked with it. Ultimately, every measurement could be evaluated in light of how recently the equipment was calibrated.

Explicitly track data “status”. Finally, it will be 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 the MPCA? Tracking this information would also allow the system to provide options for using all data or only data that has been fully verified and submitted, and would also encourage teams to upload data as soon as possible, without needing to worry about people relying on it before it has been reviewed.

5.2 Lesson 2: Catch errors early

For citizen science to be effective, it is critical that quality is understood and incorrect data identified and discarded. However, this alone does not address the sunk cost of collecting unusable data. Citizen science projects should also streamline the process for collecting and validating data as much as possible, so that errors are found early on while there is still a chance to correct them.

RQ2 showed that the current web system for River Watch is useful for *identifying errors*, especially data entry errors. However, it does little to assist in *preventing errors*, especially data collection errors. By the time data has reached the website it is usually too late to correct collection errors. The data must be simply thrown out.

What is needed is to provide feedback as soon as data is collected - while there is still an opportunity to take another measurement. Basically, the technological capabilities of an online system are needed most when volunteers are not sitting at a computer.

For River Watch, a relatively simple first step would be to automatically generate a printout that teams can take out with them when they go sampling. The printout would contain graphs with existing data for the sites they are visiting, and enough space to manually plot additional points to see where they fall in relation to the historical mean. However, a much more robust solution is detailed below.

5.2.1 Implication: Validate data in situ

The increasing ubiquity of mobile devices with Internet access should be leveraged to allow data to be uploaded to the database as soon as it is collected. 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 the case of River Watch it may still be necessary to record the data on paper for QA purposes, but this would serve as a “backup” rather than the primary input mechanism. While a mobile approach would require more tech-savvy of River Watch participants, a significant advantage the program has in this regard is the demographic doing the sampling is already quite familiar with mobile Internet technology and data-enabled smartphones.

Provide instant feedback. The real power of a mobile application would be realized by incorporating robust data validation algorithms like those implemented in the website. The key is to build on the existing ability of experienced participants to “just know” when a value is out of the expected range for a site. In addition to simple range validation, the system could compute things the coordinators know intuitively such as the expected correlation between a number of parameters related to turbidity. While this could be done via communication with a web service, the application would need to be robust enough to provide offline access to this information when the network signal was lost.

Importantly, when a data point was flagged, the team would have the opportunity to take another reading while they are still on-site. To reiterate the first lesson above, it would still be useful to ensure that process information is recorded by preserving both values and by providing an opportunity for the team to offer potential explanations for any discrepancies.

Utilize smartphone sensors. Photographs taken while on field trips would provide an additional source of rich contextual information. Like field notes, coordinators would be able to review them when evaluating data that seems to be out of range. While the location of River Watch sites is predetermined, the GPS unit in many smartphones could be utilized to auto-detect which site is being monitored.

The potential for mobile devices to serve as collectors for citizen science data has been explored by other projects. The Creek Watch mobile application[8] is designed to require no specialized equipment, relying only on observations and sensors built into the mobile device. In contrast, the application proposed here for River Watch would primarily serve to record and validate data collected through external methods. The sensors on the device would serve as input to the validation process rather than as primary data sources.

5.3 Lesson 3: Promote quality via engagement

Citizen science projects are often approached with the assumption that there is a tradeoff between collecting scientifically useful data and successfully motivating volunteers. However, this need not be the case. As RQ3 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.

Thus, when designing systems for citizen science, 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.

5.3.1 Implication: Integrate data with interpretation

Data collected through citizen science projects is often forwarded to experts for analysis. However, significant value can be added by providing interpretive tools within the system so that volunteers can explore the data themselves. The current website for River Watch provides some basic charting functionality, which could be extended to automate advanced analyses that would otherwise be too complex or time consuming for students to compute on their own. The system could also assist in generating easily-understood reports for use in school projects and outreach documents.

There are a number of existing educational resources that could be incorporated into the River Watch website. However, rather than maintaining separate “education” and “data” sections within the site, a holistic approach would be to directly link interpretive content such as lesson plans and findings with live data and graphs whenever possible.

Support social sharing. Communication tools for sharing and discussing findings will likely 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. Field notes and photographs should be given prominence in order to make the system more accessible and interesting, and to encourage teams to take good notes.

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. In general, River Watch participants have valuable domain knowledge that could be made available to others, and a social networking approach is one way to encourage it. Increased awareness of the larger community can drive home the value of consistency.

6. CONCLUSION

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 citizen science 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. This can be achieved by explicitly supporting citizen science as a form of open collaboration.

7. ACKNOWLEDGMENTS

We thank the students, teachers, coordinators and affiliates of the River Watch program for participating in and supporting this research project. We also thank anonymous reviewers who provided valuable feedback that improved the discussion in this paper.

8. REFERENCES

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

[2] B. Butler, E. Joyce, and J. Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *CHI '08*, pages 1101–1110, New York, NY, USA, 2008. ACM.

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

[4] D. Cosley, D. Frankowski, L. Terveen, and J. 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.

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

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

[7] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl. A jury of your peers: quality, experience and ownership in wikipedia. In *WikiSym '09*, pages 15:1–15:10, New York, NY, USA, 2009. ACM.

[8] S. Kim, C. Robson, T. Zimmerman, J. Pierce, and E. M. Haber. Creek watch: pairing usefulness and usability for successful citizen science. In *CHI '11*, pages 2125–2134, New York, NY, USA, 2011. ACM.

[9] 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/`.

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

[11] K. Luther, S. Counts, K. B. Stecher, A. Hoff, and P. Johns. Pathfinder: an online collaboration environment for citizen scientists. In *CHI '09*, pages 239–248, New York, NY, USA, 2009. ACM.

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

[13] 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.

[14] 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, Mar. 2011. `http://www.redlakewatershed.org/waterquality/RLWD%20SOP%20Revision%208.pdf`.

[15] D. Ribes and T. A. Finholt. Representing community: knowing users in the face of changing constituencies. In *CSCW '08*, pages 107–116, New York, NY, USA, 2008. ACM.

[16] D. 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.

[17] B. Savan, A. J. Morgan, and C. Gore. Volunteer environmental monitoring and the role of the universities: The case of citizens' environment watch. *Environmental Management*, 31:0561–0568, 2003.

[18] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Commun. ACM*, 40:103–110, May 1997.

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

[20] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282 – 2292, 2009.

[21] Y. Wand and R. Y. Wang. Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, 39:86–95, November 1996.

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

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

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