Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data

S. Andrew Sheppard
University of Minnesota &
Houston Engineering, Inc.
Minneapolis, MN, USA
sheppard@umn.edu

Andrea Wiggins
University of New Mexico &
Cornell University
Ithaca, NY, USA
andrea.wiggins@cornell.edu

Loren Terveen
GroupLens Research
University of Minnesota
Minneapolis, MN, USA
terveen@cs.umn.edu

ABSTRACT

The "real world" nature of field-based citizen science involves unique data management challenges that distinguish it from projects that involve only Internet-mediated activities. In particular, many data contribution and review practices are often accomplished "offline" via paper or general-purpose software like Excel. This can lead to integration challenges when attempting to implement project-specific ICT with full revision and provenance tracking. In this work, we explore some of the current challenges and opportunities in implementing ICT for managing volunteer monitoring data. Our two main contributions are: a general outline of the workflow tasks common to field-based data collection, and a novel data model for preserving provenance metadata that allows for ongoing data exchange between disparate technical systems and participant skill levels. We conclude with applications for other domains, such as hydrologic forecasting and crisis informatics, as well as directions for future research.

Author Keywords

volunteer monitoring; citizen science; VGI; provenance; data models; EAV; spreadsheets; mobile applications; data exchange; ICT

INTRODUCTION

Citizen science, a form of scientific collaboration that engages non-professionals in research, is an increasingly valuable means for collecting robust scientific data sets. 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. Recent high-profile citizen science projects like Galaxy Zoo [18] are largely *virtual* [32], taking advantage of the the web as infrastructure to facilitate data processing via human computation.

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 species abundance and distribution. This form of citizen science, also known as *volunteer monitoring*, has actively contributed to science and decision-making for over 100 years [4]. 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.

Virtual projects share features with the crowdsourcing style of distributed work [8], in which large contributor bases are usually required for success. In contrast, the geographic and human scale of field-based citizen science can vary widely. On one end of the spectrum are *large-scale projects* like eBird, which collects global bird abundance and distribution data [26], and the Community Collaboratorive Rain, Hail & Snow Network (CoCoRaHS) [1], a network of precipitation observers across North America. 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; the largest of several data sets includes 22 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 [34]. 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 [23] and Mountain Watch [30] 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.

Like other aspects of citizen science, the *quality of data* collected by volunteers is an ongoing area of research. Data quality is often critical to achieving project goals, and while it is generally defined as "fitness for its intended purpose" by most field research disciplines [9], its precise formulation is inherently context-dependent [23].

There are a number of mechanisms for handling data quality in citizen science [33]. Online projects can often verify results by having multiple volunteers complete the same task. However, in field-based monitoring, the data are much more individualized, containing time and location-sensitive information[1]. These 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.

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 [23], and expert review to validate the data. *Expert review* remains the most popular and broadly used mechanism for data validation in citizen science projects collecting observational data [33]; therefore, it is the primary focus of our discussion.

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.

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

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 paper, 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.

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.

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 [31, 13]. Persistent challenges in storing project data and metadata include suboptimal ICT for smaller projects [30]. 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.

Table 1. Example Citizen Science Projects

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

Given these challenges, several tools have been developed to make it easier for new projects to get started. Online systems like CitSci.org [14], Sensr [11], and CrowdMap by Ushahidi [15] 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 [7] and wq [22] 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. We have provided an implementation of the ERAV data model proposed in this work as an extension to wq[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) [3], the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata [2], and the Darwin Core [29] - 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 [5], and the conceptual W7 model [19], 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 [28], as a platform for tracking and incorporating revisions from third parties into scientific datasets.

[2]See `http://wq.io/vera` for more details.

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.

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 [12, 16], and as a way to evaluate the quality control process [6, 25]. Other efforts work to extend wikis with additional provenance tracking features, for example the Shortipedia project [27]. The geowiki model [17] 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.[3].

[3]This observation has been explored in previous work as a distinction between *descriptive* versus *experiential* volunteered geographic information [22]

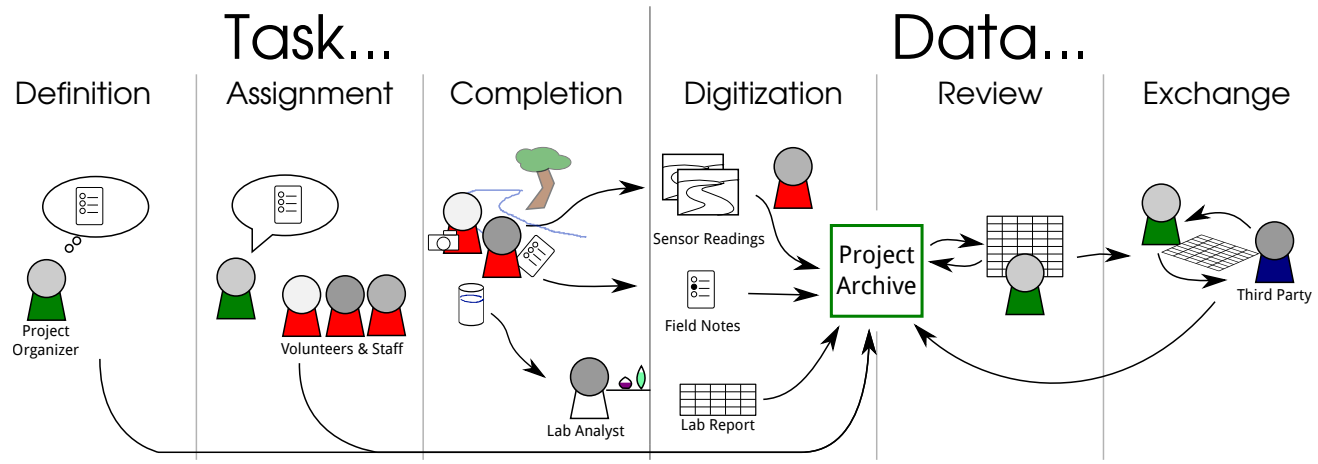

Figure 1. The field monitoring workflow. Arrows represent the flow of data and metadata between and during subtasks. (Lab analysis is relatively infrequent in citizen science projects.)

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 [21].

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 [30].

THE FIELD-BASED MONITORING 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.

Methods

The empirical data informing the inductive model development presented here were collected for two prior studies of citizen science projects [23, 30] and preliminary data from a third unpublished study. These independently conducted studies employed qualitative field research methods and shared a primary focus on *data management processes* and *ICT infrastructures* in citizen science 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 [30]. 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 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 Workflow

The process of collecting and processing field data in a citizen science project is summarized in Figure 1. We describe a general workflow, noting that each citizen science project has its own specific variations to accommodate.

Task 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 [23]. 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.

Task 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.

Task Completion

Once protocols are defined and locations assigned, it is up to the volunteers to head out into the field to collect data. This step is referred to as the *event* in our proposed data model. As noted above, this 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.

Data 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 [24]. 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 [22], 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.

Data 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 [10] – 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.

Data 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 work.

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.

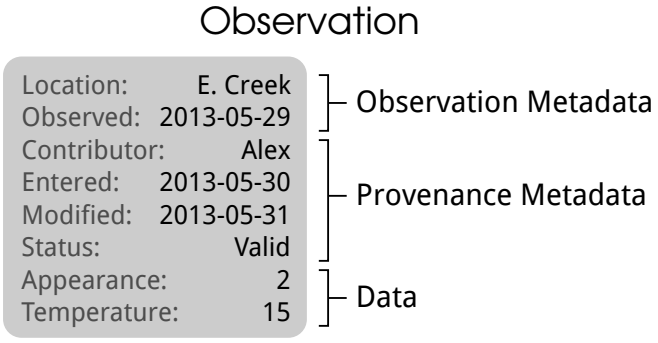

Figure 2. Example *single-table* model for representing an observation. Note that "single-table" here refers only to the structure of the observation data, as there would typically be separate Contributor and Location entities, which are represented here as text values for simplicity.

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.

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.

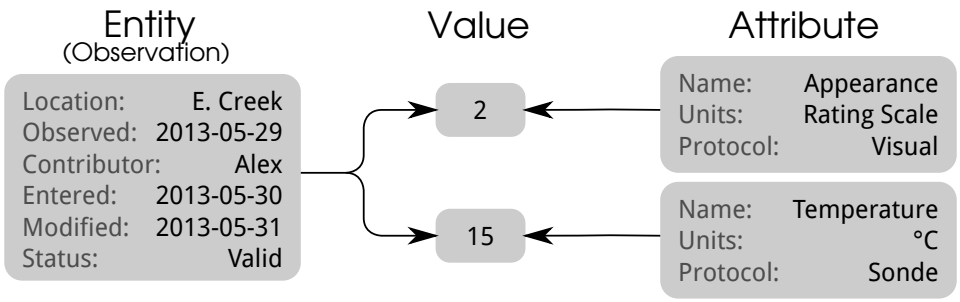

Figure 3. Simplified *Entity-Attribute-Value* model for field observations. Each rectangle represents a row in a table. Arrows represent parent-child relationships; some primary and foreign key columns are not shown.

In particular, the good old spreadsheet is how many scientific project contributors prefer to work with data [20]. 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[4]. However, mobile technology is not a panacea, and in many projects, mobile entry should not be the only contribution option. 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 [30]. Toward that end, we propose a new data model that flexibly integrates data from both bulk import and mobile field entry workflows.

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 2 shows a simple "single-table" model for storing incoming observations.

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 [22]. 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. Work is currently underway to explore the potential of applying the more flexible database model proposed in this work to CoCoRaHS.

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.

EAV: Tracking Changes to the Task Definition

The entity-attribute-value model (shown in Figure 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.

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

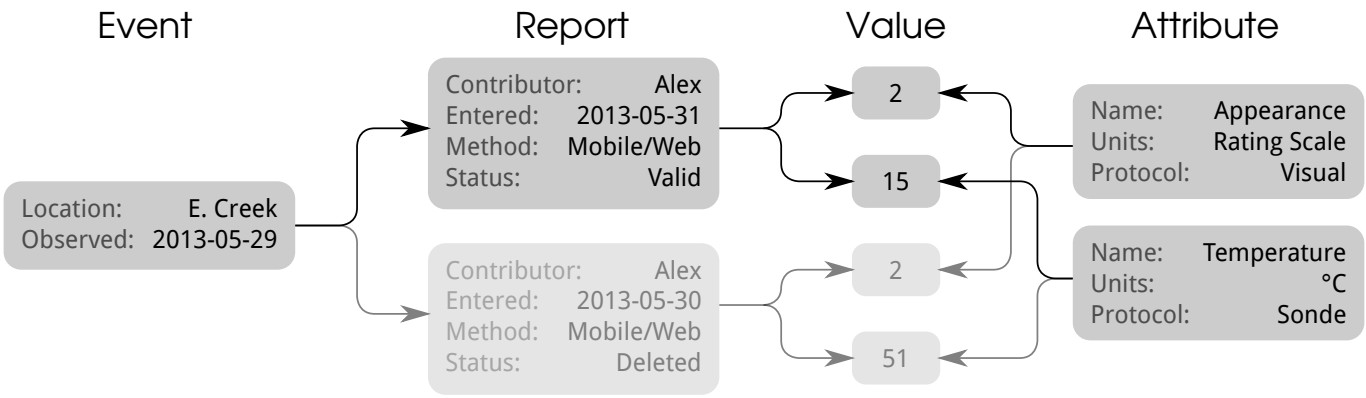

Figure 4. ***Entity-Report-Attribute-Value*** **example showing two versions of a report. In this case, all editing happened within a web interface, making it possible to automatically deprecate the old report.**

For example, eBird's internal database model includes a *Checklist* (Entity), *Species name* (Attribute), and *Observation* (Value)[5], 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 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.

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

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 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.[6] However, this assumes that revisions are done sequentially, 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* (the actual observation) from its provenance record. The *event* 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 *report* entity. *This approach enables maintaining multifaceted provenance data for the same event.*

As Figure 4 shows, whenever a report 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 report. As with EAV, the actual observations/measurements are recorded as attribute-value pairs. Importantly, these attribute-values are internally associated with each report, rather than the event.

To the casual data user and for most data analysis purposes, the event/report distinction is unimportant. For

[6]Note that "deleted" is really just a value for the *status* field; an actual SQL DELETE would cause historical data to be lost, contrary to the goals of the versioning system.

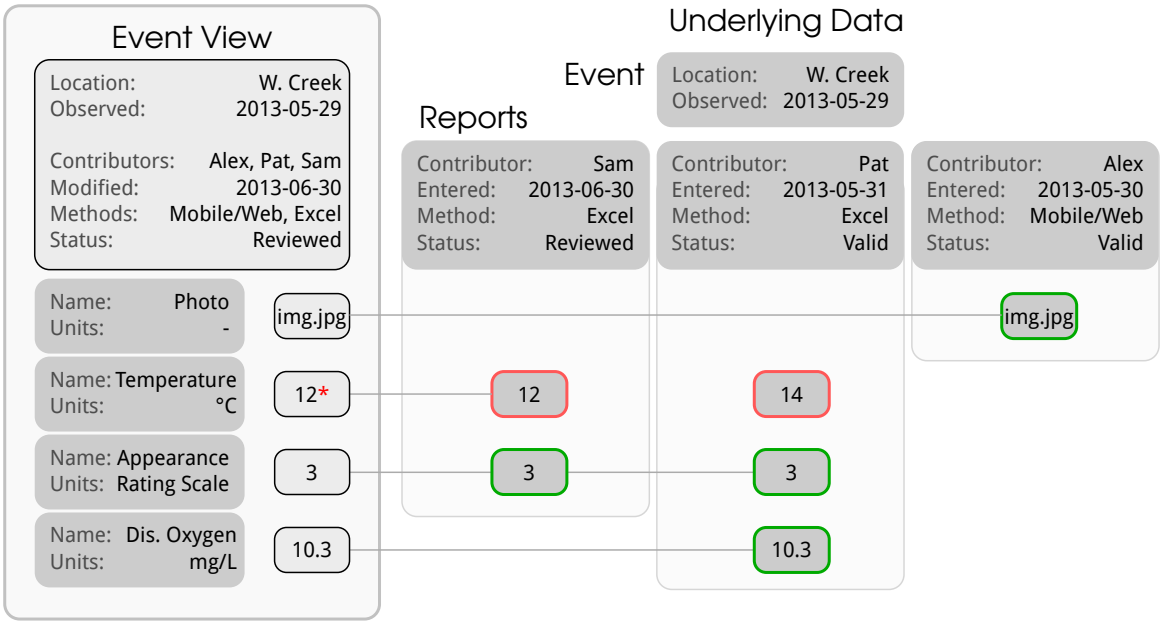

Figure 5. Resolving report information into a single "event" view. Some of the event data was imported from Excel more than once, leading to ambiguity. (For clarity, only currently active reports are shown.)

many common use cases, the attribute-values can be displayed as if they were directly associated with the event. This trick allows provenance metadata for various attribute values to be maintained internally as part of each report, while presenting a single conceptual *event* entity for analysis purposes.

"Merging" Reports & Handling Conflicts

When information about an event is created or modified outside of the system and then imported again, we cannot necessarily assume that any existing reports for the event should be deleted. For example, the incoming report may contain only supplemental data or a single additional attribute that was previously missed by mistake. Reports can also contain lab data associated with corresponding field reports, 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 reports are generated with the additional attribute-values, which are displayed together with the previously entered values as though all were directly associated with the event. This feature can be thought of as "merging" reports, though the actual data is not combined internally.

Even when the new data contains complete observation records, automatically deprecating older reports 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 reports, particularly if the incoming data set is an update to existing events 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 5 demonstrates a number of possible outcomes when multiple imported reports describe the same event. If two active reports for an event contain values for different attributes, the data is merged without issue. Even if the reports contain values for the same attributes, the data can be merged as long as the values are also the same. However, conflicting reports 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 5. By assigning a relative *authoritativeness* value to each report (e.g., the most recently *entered* report can often be treated as most authoritative), we can defer to the value from the most authoritative active report containing that attribute

whenever there is a conflict. More complex workflows might assign priorities to different report 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 Event for an Incoming Report

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

However, we cannot always control the format of batch files for upload, nor can we always ensure that updated files will contain the events' 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 event we can use to match incoming reports to events. For example, assuming only one individual is monitoring a given site once a day, we can assume that any reports for the same site and date should be associated with the same event. 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 reports will be inappropriately merged as if they were the same event, 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 event is entered with e.g. the wrong date, uploading a spreadsheet with the correct date would simply create another event 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 events upon re-import. As Figure 5 shows, this action would be instead interpreted as a partial update by the model.

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 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[7] 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, volunteer monitoring projects. In addition, we have released the source code for a generic implementation of ERAV[8] in an effort to mitigate this complexity and provide a bootstrapping platform for new projects interested in getting started with this approach.

Other Uses for ERAV

While this model was developed in the context of field-based citizen science, it has potential applications in a wide variety of domains. It may be useful in nearly any process that involves the exchange and revision of structured data between parties using incompatible software platforms or metadata practices. For example, citizen science data can be seen as a midpoint on a continuum between official or government-maintained datasets, and the more fluid crowdsourced data sets used in areas like crisis informatics.

Government agencies regularly exchange official datasets as part of their workflow. For example, the U.S. Bureau of Reclamation relies on streamflow data downloaded from the U.S. Geological Service to make water supply forecasts. In an ideal world, the data would only be used after it is fully validated and corrected. In practice, however, time constraints mean that the USBR must rely on provisional data sets that may be updated periodically without explicit information about when and why data was changed. This can make it a challenge to explain why a forecast result is coming out differently. With a model like ERAV, it will be possible to incorporate third party data and track changes to it, even if the third party is not explicitly reporting those changes.

FilteredPush [28] provides another approach to solving the problem of integrating updates from third parties into authoritative scientific datasets. However, FilteredPush requires that all involved parties use compatible software for sharing and receiving annotations to data. In contrast, ERAV can be used integrate changes from third parties not using ERAV, as long as the exchanged data has a minimally consistent structure.

[7] Concerns about performance were among the primary reasons CoCoRaHS initially chose not to utilize an EAV schema.
[8] http://wq.io/vera

Crisis informatics is another area where ERAV may prove useful. In responding to a crisis event, agencies and partners need to retrieve, organize, and evaluate data from multiple external sources (such as social networks) under rapidly changing conditions. This data must then be validated and compared against other available information. ERAV may be useful for tracking crisis reports as they move between databases and review processes. It may also prove useful in tracking revisions to public posts on social media sites that would otherwise be lost. However, the relatively unstructured nature of crisis data means that it may prove challenging to identify a usable natural key in some cases.

In summary, ERAV is most 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.

Next Steps

We plan to implement and evaluate this model with River Watch and potentially with CoCoRaHS. While we believe the model itself is sound, its success will hinge on effectively facilitating the review and integration tasks. Thus, 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 intend to explore this further in a future study. 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.

In closing, 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.

Acknowledgments

We would like to thank the participants and coordinators of the projects we studied, for their input into our understanding of the processes of citizen science. We especially thank Julian Turner for his help in relating CoCoRaHS to the concepts discussed in this paper. We also appreciate the feedback of anonymous reviewers that helped improve this paper. This work is supported in part by NSF Grant OCI-083094.

REFERENCES

1. Cifelli, R., Doesken, N., Kennedy, P., Carey, L. D., Rutledge, S. A., Gimmestad, C., and Depue, T. The community collaborative rain, hail, and snow network: Informal education for scientists and citizens. *Bulletin of the American Meteorological Society 86*, 8 (2005), 1069–1077.
2. Federal Geographic Data Committee and others. FGDC-STD-001-1998. *Content standard for digital geospatial metadata* (1998).
3. Fegraus, E. H., Andelman, S., Jones, M. B., and Schildhauer, M. 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 (2005), 158–168.
4. Firehock, K., and West, J. A brief history of volunteer biological water monitoring using macroinvertebrates. *Journal of the North American Benthological Society 14*, 1 (1995), 197–202.
5. Gil, Y., Miles, S., Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik, S. *A Primer for the PROV Provenance Model*. W3C, 2012. http://www.w3.org/TR/prov-primer/.
6. Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist 57*, 5 (2013), 664–688.
7. Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., and Borriello, G. 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*, ACM (2010), 18.
8. Howe, J. The Rise of Crowdsourcing. *Wired Magazine 14*, 6 (2006), 1–4.
9. Juran, J. M. *Quality control handbook*. McGraw-Hill, 1962.
10. Kelling, S., Yu, J., Gerbracht, J., and Wong, W.-K. Emergent filters: Automated data verification in a large-scale citizen science project. In *Proceedings of Workshops at the Seventh International Conference on eScience*, IEEE (2011), 20–27.

11. Kim, S., Mankoff, J., and Paulos, E. 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*, ACM (2013), 1453–1462.

12. Liu, J., and Ram, S. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS) 2*, 2 (2011), 11.

13. Lukyanenko, R., Parsons, J., and Wiersma, Y. Citizen science 2.0: Data management principles to harness the power of the crowd. In *Service-Oriented Perspectives in Design Science Research*. Springer, 2011, 465–473.

14. Newman, G., Graham, J., Crall, A., and Laituri, M. The art and science of multi-scale citizen science support. *Ecological Informatics 6*, 3 (2011), 217–227.

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

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

17. Priedhorsky, R., and Terveen, L. Wiki grows up: Arbitrary data models, access control, and beyond. In *Proceedings of the Seventh International Symposium on Wikis and Open Collaboration*, ACM (2011), 63–71.

18. Raddick, J., Lintott, C., Schawinski, K., Thomas, D., Nichol, R., Andreescu, D., Bamford, S., Land, K., Murray, P., Slosar, A., et al. Galaxy Zoo: An experiment in public science participation. In *Bulletin of the American Astronomical Society*, vol. 39 (2007), 892.

19. Ram, S., and Liu, J. Understanding the semantics of data provenance to support active conceptual modeling. In *Active conceptual modeling of learning*. Springer, 2007, 17–29.

20. Ribes, D., and Finholt, T. A. Representing community: Knowing users in the face of changing constituencies. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ACM (2008), 107–116.

21. Roth, M., and Tan, W.-C. Data integration and data exchange: It's really about time. In *Proceedings of the 6th Biennial Conference on Innovative Data Systems Research*, CIDR (2013).

22. Sheppard, S. A. 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*, ACM (2012), 62–69.

23. Sheppard, S. A., and Terveen, L. Quality is a verb: The operationalization of data quality in a citizen science community. In *Proceedings of the Seventh International Symposium on Wikis and Open Collaboration*, ACM (2011), 29–38.

24. Smith, A. Smartphone adoption and usage. Tech. rep., Pew Internet & American Life Project, Washington, DC, 2011.

25. Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. Information quality work organization in Wikipedia. *Journal of the American society for information science and technology 59*, 6 (2008), 983–1001.

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

27. Vrandečić, D., Ratnakar, V., Krötzsch, M., and Gil, Y. Shortipedia: Aggregating and curating Semantic Web data. *Web Semantics: Science, Services and Agents on the World Wide Web 9*, 3 (2011), 334–338.

28. Wang, Z., Dong, H., Kelly, M., Macklin, J. A., Morris, P. J., and Morris, R. A. Filtered-Push: A Map-Reduce platform for collaborative taxonomic data management. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol. 3, IEEE (2009), 731–735.

29. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS One 7*, 1 (2012), e29715.

30. Wiggins, A. Free as in puppies: Compensating for ICT constraints in citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM (2013), 1469–1480.

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

32. Wiggins, A., and Crowston, K. From conservation to crowdsourcing: A typology of citizen science. In *HICSS '11*, IEEE Computer Society (2011), 1–10.

33. Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K. Mechanisms for data quality and validation in citizen science. In *Proceedings of Workshops at the Seventh International Conference on eScience*, IEEE (2011), 14–19.

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