Free As In Puppies: Compensating for ICT Constraints in Citizen Science

Andrea Wiggins
University of New Mexico & Cornell University
159 Sapsucker Woods Rd.
Ithaca, NY 14850 USA
andrea.wiggins@cornell.edu

ABSTRACT
Citizen science is a form of collaborative research engaging the public with professional scientists. Information and communication technologies (ICT) are a leading factor in the recent spread of this phenomenon. A common assumption is that money and ICT are the ideal solutions to issues of data quality and participant engagement. The reality is instead that resource limitations often require adopting suboptimal ICT, including tools that are "free as in puppies" with hidden costs from poor usability and lack of appropriate functionality.

A comparative case study of three citizen science projects, eBird, The Great Sunflower Project, and Mountain Watch, found that projects with few ICT resources employed a broader range of strategies to address these issues than expected. The most practical and effective strategies integrated available ICT with other resources to open up new solutions and options for supporting citizen science outcomes in spite of resource limitations.

Author Keywords
Citizen science; scientific collaboration; distributed work; case study; technology-mediated participation; data quality

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces

INTRODUCTION

Citizen science involves the public with scientists in collaborative research [4]. Citizen science projects "hold out the possibility of scaling up the processes of scientific research so that they are truly global in scale and scope" [2, p. 125]. Many such projects rely upon geographically dispersed resources and contributors who work toward common goals via information and communication technologies (ICT). The dominant form of citizen science projects, found in the environmental sciences, focuses on monitoring ecosystems and wildlife populations; volunteers form a human sensor network for distributed data collection [3]. By contrast, in entirely online projects, volunteers provide data transcription and analysis services, applying basic human perception to computationally difficult image recognition tasks. The focus for this study was the more common observation-based projects traditionally called *volunteer monitoring*.

The term volunteer monitoring is now synonymous with citizen science, but is historically best known in the applied domain of natural resource management, where it has been practiced in North America for over a century [6]. Among other long-term citizen science projects, the Audubon Christmas Bird Count has been in continuous operation since 1901, with data quality and quantity reaching levels suitable for scientific research and policy decision support in the 1960's. In these projects, lay persons are trained to make scientific observations for long-term monitoring, offering new opportunities for large-scale distributed scientific research.

While citizen science cannot replace conventional science, it is already acting as a complement to professional scientific research in fields like conservation biology [4]. Citizen science projects increasingly place more emphasis on scientifically sound practices and measurable goals for public education than similar historical efforts [1]. Ample evidence has shown that under the right circumstances, citizen science can work on a massive scale and is capable of producing high quality data as well as unexpected insights and innovations [19].

Citizen science represents massive scale collaboration in science as seen nowhere else, providing an opportunity for understanding aspects of other massively distributed collaborations. In addition, many online communities and distributed collaborations face similar issues of participation and product quality, particularly in peer production settings, a common focus of research in computer-supported cooperative work research (CSCW). Prior CSCW research specific to citizen science is scant, and relatively little work focuses on the critical issues of participant recruitment, retention, and data quality, which are the focus of this study.

The remainder of this paper discusses the motivation, related work, case selection, and methodology for this study. Descriptions of the cases are followed by discussion of the use of

CSCW'13, February 23–27, 2013, San Antonio, Texas, USA.

ICT to support participation and data quality in observation-based citizen science projects and implications for practice. The conclusion identifies future directions for research related to technology-enabled citizen science.

MOTIVATION AND RELATED WORK

Despite the increasing popularity of citizen science, there are few guides to project design and implementation issues, such as selecting technologies and designing participation tasks to ensure the best possible outcomes for both the research and the participants. With the recent rise of technology-enabled citizen science, traditional citizen sciences practices are rapidly changing. In this context, successful scientific outcomes are dependent on participant recruitment, retention, and data quality, key issues facing citizen science project organizers.

Project organizers need effective mechanisms to support participant recruitment and retention, spurring research on participant motivations [15, 13]. Because the primary justification for a citizen science approach is a matter of fundamental feasibility (i.e., the research cannot be conducted in any other way, due to reliance on human cognitive competencies and/or observation over large geospatial and temporal scales), attracting large numbers of contributions from as many individuals as possible is a central goal for most project organizers.

In large-scale citizen science, the need to produce sound scientific results further conflicts with the usual conditions associated with crowdsourcing, namely that contributors are all but anonymous and their skills and knowledge base are largely unknown [8]. These challenges are compounded by the voluntary, uncompensated nature of participation: if the technologies, tasks, and goals are unappealing to would-be contributors, they simply choose not to participate. Project organizers' efforts to address these conditions are continually stymied by resource shortages endemic to volunteer work.

Participant characteristics and participation directly impact data quality. Scientists are consistently and rightfully concerned about ensuring the quality of research outputs, and public contributions come with no guarantees of expertise. Critical examination makes it apparent that the fundamental differentiating characteristic between conventional science and citizen science is the sheer number of non-professional volunteers contributing data, which has cascading effects on the research. In localized projects, typified by water quality monitoring, participation procedures may closely mimic professional protocols, with data quality addressed through extensive volunteer training and QA/QC processes [16]. For projects with a limited number of strongly motivated contributors, training volunteers in complex scientific processes can be both practical and highly successful, but this approach does not scale.

Large-scale citizen science can become truly massive: over a quarter million people have contributed classifications of images to Galaxy Zoo and related Zooniverse projects, making the contributor base orders of magnitude larger than the biggest professional scientific collaboration. In such technology-enabled projects, different approaches to participant engagement and data quality are needed. For example, using a very simple participation protocol can still yield useful data [10].

However, such simplicity is often inadequate for answering scientific research questions that demand a citizen science approach. Combined with the lack of direct oversight in distributed settings, scientists' unfamiliarity with this type of research collaboration fosters mistrust of project products despite the fact that most projects implement several strategies to ensure that the data are suitable for their intended purposes [24]. Concerns over contribution quality are the primary objections raised by critics [14], but are only a symptom of the underlying problem: researchers have yet to establish how to consistently select effective tools and mechanisms for contribution and coordination to achieve the desired scientific outcomes.

A common assumption is that money and ICT are the solution to issues of data quality and participant engagement: with more funding, superior and cheaper digital tools will yield more participation and better data quality, leading to improved scientific outcomes. The eBird project, which collects data on bird abundance and distribution, is a classic example of the basis for these assumptions. If all projects could follow suit, there would be little reason to doubt the value of citizen science.

Through comparison to Mountain Watch, a project monitoring alpine plant life cycles, and The Great Sunflower Project, which solicits observations of bees visiting sunflowers, this study revealed a different reality. Most citizen science project organizers are not well positioned to acquire adequate funding to build new ICT or improve existing tools, as their resources are consumed by project management and working to meet scientific research goals. Instead, they rely on less ideally suited technologies which may be "free as in beer" meaning that they are available at no charge [11], but are also "free as in puppies" with unexpected hidden costs of ownership. The situation faced by these projects is far more pervasive than the idealized conditions that high profile projects like eBird and Galaxy Zoo have enjoyed.

This set of assumptions and conflicting realities led to the research question for this work: *How do citizen science project organizers address key issues related to participation and data quality when resource constraints limit their ICT options?* The goal of this focus was to identify factors influencing the design and operation of citizen science projects with the greatest potential to improve project outcomes.

As the analysis presented here will show, these projects instead employed a broader range of technologies and strategies to increase participation and improve data quality than is typically assumed. Integrating this wider selection of tools and techniques with available ICT opens up new solutions and options for creating sociotechnical systems that optimize the value of citizen science for both participants and scientists in spite of resource limitations.

CASE SELECTION

A comparative case study, with theoretical sampling of three cases for in-depth study, was selected to enable a deeply contextualized understanding of the relationships between project organizers' decisions related to ICT, participation, and data quality, and their subsequent impacts. The focus of the research was on the organizers of citizen science projects rather than participants, as most prior research on citizen science focuses on participant motivations and experiences. This study therefore complements prior work and provides a foundation for integrating individual-level and group-level research for a more holistic understanding of the phenomenon.

Theoretical Sampling Criteria

Case selection included projects that are organized around primarily scientific goals and involve participants in collecting observational data about the natural world. Theoretical sampling criteria were selected from the set of project inputs identified in a conceptual framework based on literature review and prior work in small groups research [22]. The concepts of purpose, community, environment, and technologies were chosen as most likely to influence project organizers' decisions related to ICT, participation, and data quality. The application of these criteria to the selected cases are shown in Table 1.

Purpose

Two primary dimensions of purpose were considered for sampling: 1) scientific interests and goals, and 2) the mission of the project with respect to broader goals. In organizational and institutional contexts, "mission" is congruent with the guiding principles of the organization, institution, or project. Mission as used here represents explicit organizational mission or goals expressed as project mission. The missions of most observation-based citizen science projects frequently fall within the same subset of broader goals, so this dimension of the concept of purpose supports comparison. Instead, variation along the dimension of scientific interests was expected to be more important: the selected cases operate in different scientific domains, with different research goals.

Community

In citizen science, participant community is typically congruent with scientific domain variability; community as used here refers to communities of practice [20]. The intuitive expectation is that people who self-select for participation in citizen science projects focusing on birds come from a community of birders, and those who help with trailside invasive species monitoring are typically members of a hiking community.

Environment

Environment refers to organizational contexts and the broader organizational field in which a project is situated. Organizational and institutional support and constraints affect project resources and purpose. Citizen science projects are most frequently operated by academic researchers and public-sector groups, namely nonprofits and governmental agencies, so the selected cases vary across sectors. An additional related point of comparison is the resources that can be brought to the project. These include staffing and fiscal resources, summarized in full-time employee equivalents (FTEs) and approximate annual operating budget for the project.

Technologies

The technologies used for making, managing, and reporting field observations—common core tasks across the target population of citizen science projects—were one point of contrast. For example, paper-based record making in the field is very common in observational citizen science. The nature of these uses of paper, however, can differ substantially, from a protocol-based data sheet, to species lists that follow established community conventions, to free-form notes specific to individuals' field observation habits.

Two specific qualities of ICT were used in sampling, technology sophistication and data access for participants. The overall degree of technology sophistication was evaluated according to the type of web-based technologies (e.g., devoted purpose-built platforms, content management systems alone, or standard websites with data submission forms). The availability and breadth of means for data access by participants was categorized according to the extent to which contributors can interact with data. For example, very limited data access might include only a summary list of one's own observations and a static map of aggregated observation sites. Extensive data access for participants means multiple mechanisms for data access, such as: faceted data querying tools; APIs; animated and interactive maps, charts, and graphs; general audience publications; custom report generation; and/or open access to the entire data set.

Cases

The cases selected for this study are briefly introduced here; key characteristics will be discussed in more detail later.

eBird

eBird is a popular citizen science project developed by the Cornell Lab of Ornithology [18], a leading organization in the development of citizen science practice, in partnership with the National Audubon Society. Launched in 2002, eBird allows birders to report and manage birding observation records online and offers access to the full data set through a wide range of reports and data visualizations, as well as professionally curated research data products. By enticing birders to submit data through birder-centric features, eBird's data set has grown exponentially since 2005 to reach a landmark 100 million observations in August 2012, and by the start of 2013 was receiving approximately four to five million observations monthly.

eBird has generated the world's largest biodiversity data set, used not only by hobbyists and researchers, but also by land managers, policy makers, and many others. Due to its robust data validation system and enthusiastic adoption, eBird provides rich data on bird abundance and distribution at a variety of spatial and temporal scales. eBird represents a mature, well supported, and technologically sophisticated project that engages volunteers internationally on a massive scale.

Criterion		Mountain Watch	Great Sunflower	eBird
Purpose	Mission	Conservation, education, recreation	Research, education	Research, education, conservation
	Scientific Interests	Climate change effects on alpine habitats	Plant-bee relationships	Bird abundance & distribution
Intended Community		Hikers	Gardeners	Birders
Environment	Institutions	Single nonprofit	Academic	Nonprofit partnership
	Resources	1.5 FTEs, $15K	0.5 FTE, $13K	4.5 FTEs, $300K
Technologies	Paper	Structured data sheet	Structured data sheet	Variable, optional
	Digital	Organization website section	Open source CMS	Purpose-built software system
	Data access	Limited	Very limited	Extensive

Table 1. Theoretical sampling criteria in the selected cases.

The Great Sunflower Project

The Great Sunflower Project (GSP) focuses on pollinator service, that is, bee pollination activity. The project was founded in 2008 by a biology professor at San Francisco State University to collect data for her research on pollinator service, an important indicator of local ecological health. Participants grow sunflowers at homes across North America and report bee activity data online, enabling comparison of pollinator service across habitats, previously impossible at a continental scale.

While the project has been very successful in attracting volunteer interest, project leaders have struggled to turn interest into participation while also addressing pressing project sustainability concerns. The Great Sunflower Project represents a young, underfunded, and technologically disadvantaged citizen science project that has shown remarkable resilience despite substantial challenges.

Mountain Watch

Mountain Watch is a citizen science project designed and operated by the Research and Education departments of the Appalachian Mountain Club (AMC), a membership-based trail club whose mission is to support conservation, education, and recreation in the northeastern mountain ranges of the Appalachian ridge. Since 2004, Mountain Watch has enlisted hikers in evaluating air quality through visibility measurement and in collecting observations of flowering plants for climate change research. For all practical purposes, the project is geospatially constrained, collecting data primarily in the White Mountains of New Hampshire. The Whites feature the largest alpine region in the northeast U.S. with eight square miles of land above treeline, and are home to AMC's two primary visitor centers, administrative offices, and eight backcountry facilities ("High Huts").

Alpine plant monitoring, the primary focus of the project, gathers long-term data to monitor the effects of climate change on fragile alpine ecosystems by examining the timing of plant life cycle stages (phenology), such as flowering and fruiting (phenophases). Although hikers can report data online for any location in the northeastern U.S. where the target plants are found, the primary participants are hut guests. To date, hikers' data contributions show that climate change is having a stronger effect on forest plants at low elevations than high elevation alpine plants. Mountain Watch represents a mature project that has methodically fine-tuned its participation protocol over a period of several years to produce increasingly scientifically useful data.

METHODS

Field research methods were used to collect several types of data, including interviews, documents, and participant observation. As data were collected, analysis began with interview transcript coding and description of each case. Comparisons were drawn throughout the data collection and analysis process. The iterative and concurrent data collection and analysis strategy employed both inductive and deductive approaches. The remainder of this section discusses the data collection and analysis methods for this study, as well as strategies for ensuring research quality.

Data Collection

The overall case study data collection approach shared many of the characteristics of traditional ethnography, including negotiation of access, long-term participation and observation, longitudinal interviews, and field notes [17, 5]. Analytic procedures included ongoing memoing, coding, and description.

Interviews with project organizers comprised the primary data source. Interviewee selection included all project leaders and staff for each case (notably two of the projects have very limited staffing) plus representatives of diverse partner organizations both in the U.S. and internationally. Interviews also included several longitudinal interviews over a period of two to three years. In addition, hundreds of documents were used to provide background and context for the case studies, as well as to triangulate claims made in interviews.

Over 500 hours of participant observation spanning three years complemented interviews. Participant observation involved the researcher acting both as a participant contributing data for each project, and as a colleague attending citizen science organizer meetings and related events. Participating as a contributor to these citizen science projects involved hiking in the White Mountains of New Hampshire to observe alpine plants; growing sunflowers and counting bees; and learning to identify nearly 350 species of birds to report data at over 160 locations. Participation also included reading and posting to email listservs and online forums from the standpoint of a non-researcher [7].

Data Analysis

Data analysis consisted of iterative qualitative analysis guided by an existing theoretical framework [22]. The original framework was based in literature review and small groups theory, as previously mentioned, and follows an inputs-moderators-outputs-inputs structure [9], reflecting the feedback of project outputs into ongoing operation.

Throughout the course of analysis, the framework and associated schema evolved based on empirical evidence. For example, the original framework included individual processes of joining and contributing, and organizational processes of scientific research, volunteer management, and data management. The final framework included processes of science, design, organizing, and participation. These four categories encompassed prior concepts while also including other relevant subprocesses and omitting overly specific processes (e.g., joining and contributing, both part of participation) for which there was little support in the data due to the research focus on organizers.

The processes provided links between project inputs and outputs. For this analysis, the impacts on organizing and participation processes from resource constraints were evident in project-level inputs such as community, resources, institutions, and technologies. In turn, these concepts shed light on the influence of project characteristics and processes on outcomes, which included contributions, scientific knowledge, and broader impacts beyond data sets and academic publications.

The findings and interpretation were subject to review by the participants at two stages. Interview transcripts were sent to each participant for examination and verification. Key informants from each case site reviewed the case descriptions to verify factual accuracy as well as interpretation.

CASE STUDIES

The resources available in each of the cases described here varied substantially, as did their uses of ICT. This section provides background for comparison of the compensatory strategies applied when resource constraints do not permit ideal ICT-based solutions. The data contribution tasks, resources, and ICT in these projects are described below.

Data Contribution Tasks

Each of the cases collects similar data—observations about species in the natural world—but takes a different approach to participation protocols. These variations are based primarily on the expected participant base and the intended scientific applications of the data.

eBird

The basic process for contributing data to eBird is fairly simple. A birder goes birding, and makes a list of the birds observed, following one of several protocols (incidental, stationary, traveling, or area.) Along with the information about species encountered, details about the participation effort for the chosen protocol are also recorded: date, starting time, elapsed time, and number of observers, with some protocols requiring additional details about distance traveled or area surveyed. This process is only slightly more involved than traditional birding community practices, as most birders are not in the habit of collecting the effort information needed for data aggregation, nor counts of birds, records of all species seen or heard, or separate checklists for different locations. Without these details, the observations are less useful to all interested parties, including hobbyists, so eBirders often change their birding habits over time to generate more scientifically valuable data [21].

In most cases, the eBirder later enters the data through the eBird web interface or uploads data using specially-formatted email messages or Excel spreadsheets. As soon as the checklist is submitted, the birder's personal lists are automatically updated with new totals. Within 24 hours, the new observations are available through the eBird API, combined into range maps and other publicly available reports, and leaderboard rankings are updated. More recently, a third party has developed the BirdLog mobile app for data entry and submission in the field.

The Great Sunflower Project

The GSP bills its participation procedure as a four-step process, accurately describing the simplicity of the protocol. Volunteers plant Lemon Queen sunflowers (or other flowering plants added through project expansion) and can optionally report on the plant's development while they wait for their sunflowers to grow. Once the sunflowers bloom, participants choose a flower that is in the appropriate stage of development to attract bees (showing pollen) and describe the observation conditions. Next, they observe the selected bloom for fifteen minutes, recording the times at which bees visit, and optionally attempt to identify the bees. The majority of data are then entered by the volunteers into an online database, although some participants submit paper observation forms by postal mail.

Mountain Watch

Participation in Mountain Watch is largely constrained to the White Mountains. Starting at an AMC facility, hikers pick up a packet with a detailed instruction and identification guide, data sheet customized to location, and pencil in a plastic zip bag. Data sheets and reference information are also available on the Mountain Watch webpages.

Hikers locate monitoring plots using provided maps and text descriptions, and indicate which species are present and whether they are in any of the indicated phenophases (e.g., before flowering, flowering, after flowering.) The completed data sheets are then dropped off in collection boxes at any of the eight huts or the visitor centers. Observations can also be submitted online, giving contributors access to their own data, which is not currently possible for participants who submit their data on paper. While straightforward to describe, this participation protocol is quite difficult for most participants, as they are not familiar with the plant species or their phenophases.

Resources

The organizational settings were a primary factor in determining the fiscal, infrastructural, and human resources for each project. The Cornell Lab of Ornithology and Appalachian Mountain Club each had substantial but very different organizational assets that project leaders could access. For The GSP, however, institutional support was very limited; the project was just one of many faculty initiatives across a large university.

eBird

eBird's success has yielded substantial funding to support the project, with over $7 million in grant funding since the project's inception. Although this is a large figure, the grant awards were received and expended over a period of ten years, and expectations are high for a well funded project. Initially, the Lab of Ornithology provided internal venture capital funding to support project development, a particularly valuable benefit of starting the project in this unique institutional environment.

Since repaying the internal debt with grant funding from the U.S. National Science Foundation (NSF), the eBird team developed a project sustainability plan which has been recognized for its excellence as a model for sustaining digital resources [12]. The general approach is summarized very simply: "We use NSF money for innovation, and use other resources for sustainability" (Dendroica[1]). eBird's revenue sources included sponsorships, portal software licensing, endowment payouts, donations from Lab membership, and fees for location-based eBird TrailTracker kiosks at nature centers and wildlife refuges. Assuring project sustainability convinces top-level contributors and data consumers to rely on eBird as an authoritative data source and data management tool, and even more so as an increasingly longitudinal data set develops.

These income streams covered the costs of 4.25 full-time employee equivalents, including the project leaders, a web developer, department administrators, and a database administrator (all partial time with the exception of the web developer.) This represents more devoted staff time than the average citizen science project, and notably greater technical expertise. A broad skill and knowledge base was developed with lower direct investment, partly because most staff are assigned to the project part-time and partly through interorganizational partnerships. eBird is therefore able to draw upon the expertise of biologists, statisticians, computer scientists, social scientists, and complementary domain knowledge,.

The Great Sunflower Project

With less than 20 hours per week of staff time available, The GSP's most substantial ongoing challenge was the need for funding to support staffing for project coordination and communication. Early proposals for grant funding were consistently rejected due to recessionary economic conditions, so the project organizers had to explore other avenues for generating funding. Like eBird, the resultant diversification of project revenue sources is more sustainable. The GSP's new revenue streams took an entrepreneurial bent, in addition to solicitation of donations from participants.

From 2009–2011, the organizers worked with volunteers to create visually appealing calendars featuring professional quality beauty shots of bees, enhanced with additional details about each bee species and its pollination habits, emphasizing the connection between habitat conservation and food production. When calendar sales flagged, a second product was added in 2011: note cards featuring whimsical, biologically accurate illustrations of bees that further reinforce the link between bees and food.

Another strategy to bolster project sustainability was referral sales of sunflower seeds, which created an additional a revenue stream. As new revenue sources began to provide adequate funding to meet immediate operating costs, over time the organizers expressed increasing confidence in the year-to-year sustainability of the project, although the long-term outlook remained uncertain.

Mountain Watch

AMC's institutional structure and organizational resources, particularly the backcountry facilities with their constant flow of summer visitors, are a substantial asset for Mountain Watch. The eight High Huts are approximately six to eight miles apart along a 42-mile section of the 2,200-mile Appalachian Trail, operated by special permit from the U.S. Forest Service (USFS). Hut lodging capacities range from 36–90 people in bunkrooms; for a reasonable fee, guests enjoy spectacular views and hearty multi-course family-style meals.

During the summer, the huts are operated by a "hut croo"[2] of five to nine caretakers who provide hospitality, day-to-day maintenance, and emergency rescue; they are typically college students or recent graduates. The croos also include a resident naturalist with expertise or training in natural history, who makes daily educational presentations for guests on topics ranging from how Mt. Washington makes it own weather to Mountain Watch monitoring.

The internal partnership between the Research and Education departments also increased project resources, resulting in carefully designed materials for participants and Mountain Watch training for all hut croo members. The hut naturalists, who received additional training, served a special role by making regular observations at permanent plots near the huts and incorporating Mountain Watch into their presentations to hut guests. A research assistant noted, "When naturalists were giving a program about the alpine flowers, they tended to get more observations from that hut" (Clintonia).

In addition, Mountain Watch marketing materials are prominently displayed throughout the AMC facilities, with large posters in the huts and visitor centers, flyers posted on the insides of toilet stall doors, and participation materials in prominent locations. Hut croos encouraged participation in Mountain Watch at every evening presentation during dinner, at every hut, as a standard part of the croo's daily duties. Incorporating Mountain Watch into hut croo training supported ongoing participant recruitment from a constantly changing audience, as well as data validation through hut naturalists' data collection duties. Although it took considerable time to achieve, the pervasive messaging at AMC facilities demonstrated exceptional integration into the organizational culture.

[1]Latin names for species are used as pseudonyms.

[2]The term "croo" is the traditional name for these teams.

Information & Communication Technologies

As noted in the case sampling methods, the sophistication of the ICT implemented to support these projects varied substantially. The fiscal, human, and organizational resources available to each of these projects had a major impact on ICT use. While eBird had the resources to develop ICT infrastructure ideally suited to project audiences and goals, The GSP and Mountain Watch were not so fortunate.

eBird

eBird was the first citizen science project collecting data through the Internet at large scale, and online data submission is the only avenue to participation. The primary technology supporting eBird is custom online data submission and management software, developed in-house and wrapped in an open source content management system for essential communication functions, with a presentation layer for user interfaces. The unique core features include lists, alerts, and a wide variety of data reports, both personalized and general.

eBird's unique and extensive functionality is considered highly innovative by members of the birding community. During interviews, eBird was frequently identified by organizers of many other citizen science projects as having developed the most sophisticated citizen science ICT support they had encountered. The system evolved in this direction due to a consistent priority on developing functionality not available elsewhere. As an example, visualizations—particularly graphs and maps—are among these essential features because, as one organizer explained: "People like maps, and they've always liked maps, and if you can animate the maps they even like them more" (Dendroica).

The Great Sunflower Project

The Great Sunflower Project relies on the open source Drupal content management system (CMS) to support participation, although contributions are also accepted on paper data sheets. The modular system allowed organizers to pick and choose functionality to support participation, relying primarily on core features in the Drupal environment. A programmer was hired to ensure proper database management, form checking for data submissions, and usable table structures for research data retrieval. The rest of the technology development and administration is handled entirely by the project organizers. Drupal is a complex technology to learn and manage, but neither of the project's two organizers has any background in web development, a common situation in citizen science.

As the project founder quipped, open source software is "free as in puppies" for non-technologist end users, meaning that the associated costs are high when the necessary skills are not readily available. The choice to adopt an open source system kept operational costs low enough to fit a shoestring budget and permitted ongoing system upkeep without incurring substantial additional expense. The drawback, however, is that CMS like Drupal require skilled customization to function effectively for scientific data collection and need substantial work to improve usability. Many domain researchers must contract out such work because the necessary technology skills are not part of their repertoires. This challenge was exacerbated by the flat refusal of funders to provide any resources for the ICT infrastructure necessary to operate the project.

Mountain Watch

Mountain Watch contrasts sharply with the other cases in the degree to which it relies upon paper data sheets, an important non-digital technology for many field-based monitoring projects. Due to resource constraints, characteristics of the physical locations, and the natural flow of observation processes, paper remains the only practical approach for data collection at this time. For most Mountain Watch contributors participation involves *only* paper data sheets.

Online data entry and interactive map displays are a relatively new feature of the Mountain Watch webpages, which are integrated into the larger AMC website. Developed to enable participation in other northeastern forest regions and mountain ranges (e.g., the Catskill, Adirondack, and Green Mountains), this non-essential ICT provides capacity to substantially increase the scale of participation. With the exception of limited contributions from other mountain clubs (e.g., the Adirondack Mountain Club) online data entry has yet to deliver on this promise due to lack of human resources needed for outreach to promote participation outside of the Whites.

In 2009, AMC researchers also deployed plantcams, digital cameras for ongoing monitoring in outdoors environments. Most plantcams were located in established monitoring sites, so an additional benefit was that volunteers could use them to locate monitoring plots on public lands that cannot otherwise be marked due to U.S. Forest Service regulations. This technology investment was spurred by several factors: available grant funding; decreasing costs and wide availability of remote camera technologies; potential for verification of participant data; and the ability to do off-season monitoring. Some phenological events occur in late spring, when weather conditions are so severe that no one will venture out to make observations.

DISCUSSION

Differences in project resources profoundly impacted the use of ICT to support participant recruitment, retention, and data quality in each of these cases. As was expected, monetary and human resources had the most direct influences on project design and outcomes, but resources unique to organizational context were also a substantial and often overlooked factor. In the discussion of participant recruitment, retention, and data quality that follows, eBird provides an example of the expected uses of ICT tailored to the project, while The GSP and Mountain Watch demonstrate strategies that compensated for ICT limitations.

Recruiting Participants

Recruitment and retention of participants is a common concern for citizen science projects that need to accumulate a large data set. In many respects, all three projects used the same tried-and-true volunteer recruitment strategies employed in many other settings. Across the cases, ICT was not the most effective tool for recruitment. In theory, being open to web-based contributions would make project participation

more visible to a broader audience and ease recruitment efforts. This was only partially true; contributors were recruited most effectively through direct in-person contact.

The three eBird project leaders estimated that they each spent approximately 100 days per year traveling to make presentations as keynote speakers at bird festivals and reach out to local groups such as bird clubs and Audubon chapters. Outside of the U.S., these efforts are extended by staff of partner organizations, such as Bird Studies Canada, who engage in similar outreach in their own communities. By contrast, The GSP had great success in initial recruitment by requesting only a few individuals email information about the project to their social networks. The level of follow-through from these initial recruits, however, was very low. Partnering with similarly focused organizations and hiring an Outreach Director to promote local participation generated better recruitment results, as did newspaper and magazine articles featuring The GSP. Mountain Watch organizers were able to employ a different strategy altogether, exploiting location-based organizational resources for ongoing recruitment of a high turnover volunteer base. ICT for online data submission, however, had yet to pay off in increased recruitment and participation.

Unlike the other two projects, eBird was able to create ICT-based features that further improved participant recruitment by leveraging existing birder community practices. eBird organizers did not attempt to displace the well established birding listservs through which individuals share sightings with their local communities. Instead, they created features to work with the existing community infrastructure: eBirders can have nicely formatted checklists emailed to them to forward to friends and local listservs. These messages are more easily produced, readable, and useful than the average trip report and include footer text that reads, "This report was generated automatically by eBird v3," a subtle form of social recommendation. Interviewees reported that in some areas the majority of messages on the email lists are forwarded eBird reports. The well established social structure of birding communities made emailed checklists an effective way to take advantage of participants' social networks to effortlessly extend recruitment.

Due to differences in community structures and the lack of ICT tools to accomplish similar ends, however, neither of the other two projects were able to mimic eBird's strategy. Instead, both The GSP and Mountain Watch used paper-based data submission as a means of expanding the contributor base and increasing data submissions. For Mountain Watch, this was a matter of practicality on several levels. Collecting paper data sheets at the AMC facilities yielded a higher return rate because participants did not have to go out of their way to submit data. Expecting hikers to enter this data online after returning home from an exhausting vacation in the Whites would be unrealistic: the records become lost, mangled, or forgotten, among other reasons for neglecting to submit data.

The GSP, on the other hand, emphasizes online data entry but accepts paper data sheets by postal mail. In addition to the practical considerations of outdoors data collection, paper data sheets are one way that The GSP organizers try to be inclusive of "die hard retirees that want to participate, that want to feel part of the project, but yet are really still very uncomfortable with online entry, who get confused by websites" (Bombus). The level of difficulty that older adult participants encountered with online data submission was a substantial barrier for some individuals, due in no small part to the lack of resources to devote to improving the usability of the Drupal site. The use of paper data sheets helped make up for the organizers' inability to make the ICT more accessible by expanding the means of participation.

While all three projects used many similar recruitment techniques from established volunteer management practices, each found that in-person recruitment was most effective. For eBird, ICT-based marketing was effective when spread by participants through their local social networks via emailed and shared checklists that dovetail with existing community practices, benefit participants, and advertise eBird functionality. For The GSP and Mountain Watch, however, developing such features was out of the question. Paper data sheets—a more accessible technology for some participants—and refinements to the participation protocols both helped compensate for this limitation in The GSP, and in Mountain Watch, saturation marketing throughout physical facilities and daily presentations to hut guests were highly effective.

Participant Retention

In most cases, retaining participants is nearly as high priority as recruiting them because ongoing participation is expected to yield more and better data submissions. Unfortunately, the ICT features that clearly supported retention for eBird were among the most expensive to develop, making them unrealistic options for The GSP and Mountain Watch.

eBird developers spent years developing rapid feedback in the form of birder-centric reports and visualizations that reward contributors for submitting data by providing access to information they want. These included leaderboards that built on long-standing traditions of friendly competition in the birding community, subtly incorporating social rewards based on community norms. Participants find the data displays very satisfying for their personal interests in bird data and for elevating their social standing in the birding community through visibility of their contributions.

While eBird's simple bar charts and ranked reports may seem an easy solution, they required extensive development time, and project staff remarked upon the continual difficulty of integrating custom software with the CMS and presentation layers that comprise the full system. The latest and greatest of these visualizations, animated migration maps, required millions of data points, invention of new spatiotemporal statistical modeling techniques, and high performance grid computing to generate. In short, the innovative and popular data visualizations cost a lot of money.

Interviewees connected these high quality visualizations with several benefits to participants that can also support retention. According to a project organizer, "the fact is that to see the dynamics, spatially and temporally, of how these things change, and to do it at such a broad scale as we can do it, is

transformative in the way people think about biodiversity and natural history" (Dendroica). For many participants, the data visualizations may be just another way to find the information they desire, but for others, seeing the aggregated data in a different way can stimulate a change in the way they understand relationships between biodiversity and habitat preservation, reinforcing the value of participation.

These ICT-supported offerings, while in many ways ideal for a distributed contributor base, are practically impossible to implement for small underfunded citizen science projects. The Mountain Watch website now includes an interactive map of phenology observations, although the relatively low emphasis on ICT use for the primary contributor base makes the utility of this visualization for contributor retention questionable, and it may instead be more valuable for recruitment. In addition, most Mountain Watch participants are one-time contributors on a brief vacation in the mountains, and are therefore less invested in submitting more data or making use of data visualization tools. Instead of focusing on retention, the organizers therefore prioritized recruitment.

The GSP provided a static map on the project website showing locations of bee sightings, which took three years to produce. The alternate solutions the project organizers implemented to improve retention included modifying the observation protocol to make sampling times shorter (from 30 to 15 minutes) and accepting data for a wider variety of common garden flowers beyond sunflowers, in response to popular demand. These changes better supported participation by families with small children and those whose sunflowers fail, which reduced participation in the first few years due to bad seeds and unfavorable weather conditions.

Adding more monitoring species also encouraged additional contributions by volunteers whose gardens already contained multiple target species. Further, it will eventually enable new research by accumulating comparative data of bee visitation rates across plant species, which was not an initial goal but will make an additional novel and valuable scientific contribution. The GSP organizers also experimented with other strategies to support retention, including different modes of communication with participants and explicit rewards, such as sending sunflower seeds to participants who have submitted multiple data points for several years to encourage them to continue.

While project organizers often expressed yearning to provide reports and visualizations similar to those eBird features, The GSP and Mountain Watch organizers all felt that this was a pipe dream. The cost of ICT tailored to support specific project needs and goals was simply too high for their limited resources. Instead, they found other ways to support retention by expanding options available to participants, trying new communication strategies, or simply focusing instead on recruitment.

Maximizing Data Quality

Contribution quality is a high priority in the context of scientific research. Carefully structured participation tasks are a very important facet of quality assurance. Data verification is also needed, however, and ICT are among the few truly scalable solutions. Tools for expert data review, a favored strategy, usually require custom development that substantially increases the cost of scalable quality control.

Not surprisingly, the data quality strategies in these cases were differentiated primarily based on use of human resources versus reliance on ICT. eBird was again an exemplar, with several features of participation protocols to improve incoming data supported by rigorous, scalable data verification. This process employs both customized ICT and expert birders to generate scientifically reliable data. Notably, eBird was uniquely able to implement a set of observation protocol that mirrored existing birding practices and therefore have required no changes over time [22]. eBird's other data quality mechanisms are largely based on customized ICT.

First, the eBird data entry interfaces present predictive checklists that reinforce careful species identification by displaying only the most likely species for the specific place and time of year, based on both expert input and existing data. Second, eBird uses a custom taxonomy with less specific taxonomic categories for birds that are difficult to identify in the field. A project leader described the importance of the specialized taxonomy to data quality: "We are not going to do any great science with [those data], but we are going to keep someone from just pigeonholing a bird because they think that's what it is but they're not really sure" (Stercorarius). This is a valuable strategy, given the uncertainties of field observation and variability in observer skill, but required substantial effort to create.

Finally, data review by experts plays a crucial role in quality control. eBird's reviewer network includes approximately 500 hand-selected individuals in North America, plus a few international reviewers. Volunteers performing this role often consult historic records for the location, ply personal local knowledge, and engage data contributors in email exchanges. After reviewing algorithmically flagged records, the reviewers render a judgment of whether the sighting is valid or not, which is then added to the record; observations that are considered invalid still appear in the user's life list but are not added to the research data set.

Needless to say, the specialized review tools and network of experts that made eBird's system scalable and highly reliable are also out of reach for projects with less funding for ICT and personnel. However, as contribution rates continued to increase sharply, the stress on the reviewer system had started to limit project capacity, leading to development of increasingly sophisticated ICT functionality [23]. Instead, The GSP organizers spent two years ironing out the participation protocol, during which time most data could not be salvaged, and then integrated authoritative data sets (housing density data) upon discovering that participants' characterizations of their garden locations were too subjective to be reliable.

The GSP also invested limited funds in database design, according to the project founder: "When I spent money [on ICT], it's really been to manage the forms where data gets entered, because I want to make sure that there's no way I

can screw it up" (Apis). As with other projects using online data submission forms, implementing these measures improved data quality at an additional but modest expense that was substantially outweighed by the benefit.

Mountain Watch responded to similar participation process issues with an in-depth study of factors affecting data quality, and used multiple sources of complementary data. The organizers based protocol modification on rigorous scientific research, complete with statistical analyses that conclusively identified problems related to task complexity.

Evaluation of the original protocol highlighted participants' difficulties with location description and identification tasks. These insights led organizers to shift from open-ended locations to specific monitoring plots on a map, with detailed descriptions for reference. They also added lists of the known monitoring species occurring at each of the plots, reducing uncertainty in species identification, and tested the new data sheet the following year to verify improved data quality.

> People hiking from Mizpah [Spring Hut] to Lakes [of the Clouds Hut] were providing better data than the people hiking from Lakes to Mizpah because [they] had the new data sheet and were going to specific places. People hiking from the other direction but on the same trail, they just didn't have the focus, and their data was kind of all over the place. (Clintonia)

Developing a viable protocol was a substantial human resource investment for Mountain Watch, but the experimental design for evaluation of data quality paid off with certainty of the benefits of making further changes. The statistical analyses that capitalized on the scientists' research skills were a more rigorous approach to appraisal of the data collection instrument than is typically seen in citizen science projects.

Another strategy to ensure scientific data quality was integration with conventional scientific research and use of complementary technologies. Mountain Watch organizers implemented three different methods to capture data about phenophases: citizen science volunteers, trained hut naturalists, and automated data collection using plantcams. These data sources also complemented ongoing data collection by scientists, as some of the observation sites were at existing research plots.

As previously mentioned, the installation of plantcams served multiple purposes, including potential application to verifying volunteers' data. Actually comparing plantcam data to hikers' observations, however, is a more complex undertaking than it might initially appear. Processing image data from plantcams for this type of phenology monitoring requires human effort. This is a common focus for analysis-oriented citizen science projects like Galaxy Zoo, but requires customized technological infrastructure that is not readily available to most citizen science projects and too costly to customize to their purposes.

While eBird used ICT to develop a scalable, reliable approach to data verification, such purpose-built technologies are not available for most citizen science projects. The GSP instead invested time in protocol development and data integration and spent some limited funds on database design. Mountain Watch, on the other hand, was able to take advantage of internal research skills and low-cost labor for a rigorous multi-season study that honed not only the participation protocol, but also the data collection instrument. Using multi-layered data collection to supplement volunteers' data with complementary data from scientists, trained naturalists, and plantcam images was another promising strategy for ensuring data quality.

Implications

Returning to the concepts from the theoretical framework, these cases verified the expected links between project inputs and outputs through processes of design, organizing, contribution, and science. The cases also demonstrated that these factors and processes are strongly intertwined: there is no single input that guarantees data quality, nor can any one process effectively support participant recruitment and retention. As a complex sociotechnical phenomenon, developing ICT for technology-supported citizen science requires a holistic understanding of project goals, practices, and resources.

These findings challenge the notion that better ICT is the best solution to participant recruitment, retention, and data quality in citizen science. The reality is instead that many projects' resource limitations require adopting suboptimal ICT, including tools that are "free as in puppies" with hidden costs in the form of poor usability and lack of appropriate functionality. The primary insight that emerged was a broader view of potential solutions beyond ICT, demonstrating that the most practical and effective strategies under these constraints involved combining ICT with other resources, particularly human expertise.

The implications of these findings suggest that ICT selection and development should focus primarily on project goals, known characteristics of the participant audience that influence recruitment and retention, and data quality requirements for scientific outcomes. Tempering the instinct to rely primarily on ICT to address these concerns, it is likewise important to identify ICT-related constraints and work to find compensatory strategies based on other available resources, which are sometimes taken for granted by organizers and overlooked by technologists.

CONCLUSION

For citizen science projects with few resources, managing ICT can be burdensome due to low levels of technology expertise among project organizers. In most cases, technologies with the least complexity and lowest cost are the only sustainable choices despite the fact that nearly all organizers are aware that far better results could be obtained with more sophisticated systems. The affordability of tailoring ICT to project needs and audiences, however, remains a constant limitation.

As the case studies showed, however, alternate strategies that rely less on ICT and instead leverage other types of resources can provide effective ways of compensating for less than ideal ICT support. Direct contact with potential participants was

the most successful recruitment strategy, and although it can be bolstered with ICT functionality, this turned out to be a relatively low priority for ICT use. Participant retention was more challenging when organizers were unable to provide engaging data displays, but modifications to participation activities and communication strategies were useful techniques for improving retention. Those citizen science projects with adequate resources have successfully formulated robust, scalable systems to support the scientific rigor of the research [25], and while these tools are currently out of reach for most smaller projects, turning to other resources can achieve similar data quality for a wide range of citizen science projects.

Similarities between citizen science and online communities, distributed scientific collaboration, and peer production suggest opportunities for future work examining how findings from CSCW studies of massively distributed collaborations may generalize across contexts. Identifying means of improving engagement and product quality that do not rely solely on ICT provides a reminder for CSCW researchers that technologies are not a solution in and of themselves. Ideal technologies are not always available to practitioners, particularly in resource-constrained contexts, and remaining open to complementary strategies to support distributed collaboration can provide viable solutions. Case studies such as those presented here can also provide a frame of reference to both practitioners and researchers exploring alternate approaches for improving upon large-scale open collaboration.

As ICT supporting citizen science continues to evolve, the emergence of cyberinfrastructures and platforms to support projects faced with resource limitations will likely begin to address these issues more comprehensively. In the meantime, attention to the context of practical constraints surrounding ICT use, incorporating creative strategies that do not rely entirely on ICT, and leveraging small investments in ICT for maximum benefit are viable approaches to improving citizen science project outcomes.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grants VOSS-0943049, SOCS-0968470, and OCI-0830944. I am deeply grateful to the organizers of eBird, The Great Sunflower Project, Mountain Watch, and their colleagues and partners who shared their time and experiences for this study. Finally, I thank the AC and anonymous reviewers, whose feedback substantially improved the work.

REFERENCES

1. Bonney, R., Cooper, C., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K., and Shirk, J. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience 59*, 11 (2009), 977–984.
2. Bowker, G. *Memory Practices in the Sciences*. MIT Press, Cambridge, MA, 2005.
3. Cohn, J. Citizen science: Can volunteers do real research? *BioScience 58*, 3 (March 2008), 192–107.
4. Cooper, C., Dickinson, J., Phillips, T., and Bonney, R. Citizen science as a tool for conservation in residential ecosystems. *Ecology and Society 12*, 2 (2007).
5. Emerson, R., Fretz, R., and Shaw, L. *Writing Ethnographic Fieldnotes*. University of Chicago Press, Chicago, IL, 1995.
6. 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.
7. Hine, C. *Virtual Ethnography*. Sage Publications Ltd, Thousand Oaks, CA, 2000.
8. Howe, J. The Rise of Crowdsourcing. *Wired Magazine 14*, 6 (2006), 1–4.
9. Ilgen, D., Hollenbeck, J., Johnson, M., and Jundt, D. Teams in organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology 56* (2005), 517–543.
10. Kim, S., Robson, C., Zimmerman, T., Pierce, J., and Haber, E. Creek watch: Pairing usefulness and usability for successful citizen science. In *Proceedings of the 2011 International Conference on Human Factors in Computing Systems*, ACM (New York, NY, 2011), 2125–2134.
11. Klang, M. Free software and open source: The freedom debate and its consequences. *First Monday 10*, 3-7 (2005).
12. Maron, N., Smith, K., and Loy, M. Sustaining Digital Resources: An On-the-Ground View of Projects Today. Tech. rep., Ithaka S+R, New York, NY, 2009.
13. Nov, O., Arazy, O., and Anderson, D. Dusting for science: Motivation and participation of digital citizen science volunteers. In *Proceedings of iConference 2011*, ACM (New York, NY, 2011).
14. Roman, D. Crowdsourcing and the question of expertise. *Communications of the ACM 52*, 12 (2009), 12.
15. Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., and Jacobs, D. Dynamic changes in motivation in collaborative ecological citizen science projects. In *Proceedings of the ACM 2012 conference on Computer supported cooperative work*, ACM (New York, NY, 2012).
16. Sheppard, S., 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.
17. Spradley, J. *The Ethnographic Interview*. Wadsworth Publishing Company, 1979.
18. Sullivan, B., Wood, C., Iliff, M., Bonney, R., Fink, D., and Kelling, S. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation 142*, 10 (2009), 2282–2292.

19. Trumbull, D., Bonney, R., Bascom, D., and Cabral, A. Thinking scientifically during participation in a citizen-science project. *Science Education 84*, 2 (2000), 265–275.

20. Wenger, E. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1999.

21. Wiggins, A. eBirding: Technology adoption and the transformation of leisure into science. In *Proceedings of iConference 2011* (Seattle, WA, 2011).

22. Wiggins, A., and Crowston, K. Developing a conceptual model of virtual organisations for citizen science. *International Journal of Organisational Design and Engineering 1* (2010), 148–162.

23. Wiggins, A., Gerbracht, J., Lagoze, C., Yu, J., Wong, W., and Kelling, S. Crowdsourcing citizen science data quality with a human-computer learning network. In *Proceedings of the Workshop on Human Computation for Science and Computational Sustainability*, Neural Information Processing Systems Foundation (Lake Tahoe, NV, 2012).

24. Wiggins, A., Newman, G., Stevenson, R., 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).

25. Wood, C., Sullivan, B., Iliff, M., Fink, D., and Kelling, S. eBird: Engaging birders in science and conservation. *PLoS Biology 9*, 12 (12 2011).