Journal of Database Management
Volume 29 • Issue 1 • January-March 2018

Beyond Micro-Tasks:
Research Opportunities in Observational Crowdsourcing

Roman Lukyanenko, University of Saskatchewan, Saskatoon, Canada

Jeffrey Parsons, Memorial University of Newfoundland, St. John's, Canada

ABSTRACT

The emergence of crowdsourcing as an important mode of information production has attracted increasing research attention. In this article, the authors review crowdsourcing research in the data management field. Most research in this domain can be termed tasked-based, focusing on micro-tasks that exploit scale and redundancy in crowds. The authors' review points to another important type of crowdsourcing – which they term observational – that can expand the scope of extant crowdsourcing data management research. Observational crowdsourcing consists of projects that harness human sensory ability to support long-term data acquisition. The authors consider the challenges in this domain, review approaches to data management for crowdsourcing, and suggest directions for future research that bridges the gaps between the two research streams.

KEYWORDS

Citizen Science, Conceptual Models, Crowdsourcing, Data Management, Data Models, Observational Crowdsourcing

INTRODUCTION

Recent years have seen a major shift in knowledge production via crowdsourcing, wherein increasingly work is being done by distributed members of the general public (the crowd), rather than employees or traditional subsidiaries. Crowdsourcing promises to dramatically expand organizational computing power and "sensor" networks, making it possible to engage ordinary people in large-scale data collection (Brabham, 2013; Doan, Ramakrishnan, & Halevy, 2011; Franklin, Kossmann, Kraska, Ramesh, & Xin, 2011; Garcia-Molina, Joglekar, Marcus, Parameswaran, & Verroios, 2016; Li, Wang, Zheng, & Franklin, 2016).

Applications of crowdsourcing are rapidly expanding and power such diverse activities as corporate product development, marketing, public policy, scientific research, graphic design, software development, and writing and editing. Crowdsourcing is increasingly tasked with tackling difficult societal and technological challenges, such as climate change (Theobald et al., 2015), natural disasters (Brabham, 2013) and commonsense reasoning in artificial intelligence (Davis & Marcus, 2015).

Organizations integrate crowdsourcing into internal decision making and operations. Fortune 500 companies maintain digital platforms to monitor what potential customers are saying and understand customer reactions to products and services. They also use consumer feedback to design better products and monitor market changes (Abbasi, Chen, & Salem, 2008; Barwise & Meehan, 2010; Brynjolfsson & McAfee, 2014; Delort, Arunasalam, & Paris, 2011).

DOI: 10.4018/JDM.2018010101

Crowdsourcing turns problem-solving capacity and data into commodities, making them available on-demand. For example, many municipalities in the United States now subscribe to CitySourced.com, which harnesses citizens' reports of crime, graffiti, potholes, broken street lights, and other civic issues, to better support infrastructure management. In a more general setting, Amazon's Mechanical Turk (mturk.com), CrowdFlower.com, and Clickworker.com maintain pools of "crowdworkers" that companies hire on-demand to perform small problem-solving tasks.

There is also a proliferation of platforms for automatically generating data collection forms that can be easily configured and rapidly launched on a large scale. Projects such as EpiCollect.net, SciStarter.com, or SmartCitizen.me make crowdsourcing possible for organizations and even individuals, requiring little technical expertise and infrastructure. Crowd-powered extensions of word processors, such as Soylent, enlist crowds for document writing and editing (Bernstein et al., 2015). In addition to becoming a mainstream commercial service, crowdsourcing has become a major resource for scientific research (Goodman & Paolacci, 2017).

Crowdsourcing presents several data management challenges. Unlike traditional data collection in organizations, in crowdsourcing there are typically weaker constraints on who can participate. This creates the challenge of managing data produced by often anonymous users with varying levels of domain expertise or motivation (Lukyanenko, Parsons, Wiersma, Sieber, & Maddah, 2016). Furthermore, in many projects participation is voluntary, making it difficult to engage users in eliciting information requirements (e.g., to guide database design) or in improving the quality of existing data (e.g., to clarify a particular data entry, or request additional information) (Chen, Xu, & Whinston, 2011). These challenges offer exciting opportunities for data management researchers to design innovative solutions.

In this paper, we survey current research on data management in crowdsourcing. We argue that two major, largely disconnected, streams of crowdsourcing research exist: (1) studies that investigate uses for crowds as links in a larger technological chain and tend to focus on small, granular and well-defined micro-tasks (task-based crowdsourcing); (2) studies that explore the potential of crowds as "sensors" in the environment that can be leveraged in organizational decision making and innovation (observational crowdsourcing). Until now, the data management community has studied primarily the first kind of crowdsourcing, resulting in a serious gap in resolving challenges associated with the latter type. With this in mind, we analyze the challenges and opportunities associated with observational crowdsourcing environments and suggest directions for future database research in this context.

TYPES OF CROWDSOURCING

Driven by a combination of: (1) sustained interest from organizations in harnessing the knowledge of ordinary people; and (2) formidable research challenges, the data management community has increasingly adopted crowdsourcing as a serious research topic. The growth of crowdsourcing data management research is evidenced by the proliferation of panels, workshops, and special journal issues on this topic. Examples include: *VLDB* panels and workshops starting from 2007's "Web 2.0 and Databases"; *ACM SIGMOD 2009*'s panel "Crowd, Clouds, and Algorithms"; *ACM CIKM* "CrowdSense" 2012 and 2014 workshops; *ACM SIGKDD* workshops; *AAAI Conference on Human Computation & Crowdsourcing* (*HCOMP*) since 2009; and a growing number of journal articles.

The importance of research on crowdsourcing data management is underscored by its relevance to other major data management problems. Crowdsourcing is a key contributor to the growth of user-generated content and is commonly considered together with social media and social networking (Faraj, Jarvenpaa, & Majchrzak, 2011; Germonprez & Hovorka, 2013; Kane, Alavi, Labianca, & Borgatti, 2014; Levina & Arriaga, 2014; Zwass, 2010). As crowds, due to their scale, are capable of generating massive volumes of rich and heterogeneous data, crowdsourcing exemplifies the "big data" management challenge (Carlo Batini, Rula, Scannapieco, & Viscusi, 2015; Jagadish et al., 2014). As crowd data is often sparse, and of uncertain quality, crowdsourcing research is relevant to work on noSQL databases, integration of heterogeneous data sources, and data quality (Cattell, 2011; Kosmala, Wiggins, Swanson, & Simmons, 2016; Lukyanenko, Parsons, & Wiersma, 2014b). Finally, tapping into human problem-solving ability, crowdsourcing promises to complement technological shortcomings of machines (e.g., through human-powered query execution, sorts, joins, or inference-making) (Bernstein et al., 2015; Davis & Marcus, 2015; Franklin et al., 2011).

Despite such potential, the scope of data management research in crowdsourcing has been narrow. Based on current publications in information systems, computer science, and software engineering, we focus on two clearly identifiable streams of crowdsourcing research – the dominant (small) task-based crowdsourcing and the less common observational crowdsourcing – that, while related, are pursued largely independent of each other.

On one hand, there is growing body of work that aims to leverage crowds as links in a larger technological chain, in which people are seen primarily as problem-solvers in small, granular and well-defined tasks (Amsterdamer & Milo, 2015; Garcia-Molina et al., 2016; Kittur, Chi, & Suh, 2008; Li et al., 2016; Paolacci, Chandler, & Ipeirotis, 2010; Weld, 2015). In this context, Amsterdamer et al. (2015) liken crowds to “an external (and very slow, potentially unreliable) hard drive” that can be queried on demand for information. Since this work typically considers crowds as useful for performing small, independent, and well-defined tasks, we term this stream task-based crowdsourcing.

On the other hand, there is growing recognition of the potential of crowds to support organizational decision-making, operations and innovation through the ability to observe and report on their environment (Brabham, 2013; Brynjolfsson & McAfee, 2014). With an explicit organizational focus, this line of research aims to better understand the advantages and limitations of crowdsourcing as a unique form of organizational “sense-making” (Brabham, 2013; Brynjolfsson & McAfee, 2014). Researchers in this stream investigate a variety of issues, including motivation of crowds (Coleman, Georgiadou, & Labonte, 2009; Daugherty, Eastin, & Bright, 2008; Nov, Arazy, & Anderson, 2011; Prestopnik & Crowston, 2011), data quality in the crowds (Ogunseye, Parsons, & Lukyanenko, 2017; Parsons, Lukyanenko, & Wiersma, 2011; Sheng, Provost, & Ipeirotis, 2008; Sheppard, Wiggins, & Terveen, 2014; Wiggins, Newman, Stevenson, & Crowston, 2011), organization of crowds (Arazy, Nov, Patterson, & Yeo, 2011), design of crowdsourcing platforms (Lukyanenko, Parsons, & Wiersma, 2011; Prestopnik & Crowston, 2011; Wiggins et al., 2013), and utilization of crowd-produced resources in organizations (Majchrzak & More, 2011). We focus on those issues relevant to data management. As organizational data collection is typically a continuous, on-going process that involves observing or sensing the broader enviornment, we term this research stream ***observational crowdsourcing***.

Note that other types of applications of crowdsourcing exist, including question-answering sites, contests, reviews, and fundraising (Brabham, 2013; Dissanayake, Zhang, & Gu, 2015; Doan et al., 2011; Leimeister, Huber, Bretschneider, & Krcmar, 2009). We limit our focus to the two streams of research that have explicit data management implications (but as databases underlie most crowdsourcing technologies, future research should consider other types of crowdsourcing as well).

TASK-BASED VS OBSERVATIONAL CROWDSOURCING

In task-based crowdsourcing, crowds are data providers and problem solvers in a larger technological chain. Characteristic features of this type of crowdsourcing include:

1. Small, well-defined tasks;
2. Activities take place primarily online;
3. Crowdworkers are typically (but not always) paid.

Major examples of task-based crowdsourcing platforms include Amazon Mechanical Turk, CrowdFlower, and Microworkers. Using such platforms, designers can pose tasks to a crowd of workers for a small payment. Typical small (“micro”) tasks include classifying items, providing missing values, sorting and filtering items, and comparing items; however, more complex tasks, such as answering open-ended questions, are possible (Amsterdamer et al., 2015; Amsterdamer & Milo, 2015; Kittur et al., 2008; Paolacci et al., 2010; Sorokin & Forsyth, 2008). These tasks can be sent to crowd database engines that combine traditional SQL statements with user-defined functions (Franklin et al., 2011; Park et al., 2013). Micro-task data management projects include CrowdDB (Franklin et al., 2011), sCOOP and Deco (Parameswaran, Park, Garcia-Molina, Polyzotis & Widom, 2012; Park et al., 2013), Qurk (Marcus, Wu, Karger, Madden, & Miller, 2011b) and CrowdFill (Park & Widom, 2014).

Task-based crowdsourcing also includes some citizen science projects such as GalaxyZoo, Snapshot Serengeti, and FoldIt (Cardamone et al., 2009; Clery, 2011; Simpson, Page, & De Roure, 2014). Unlike Amazon Mechanical Turk, these projects do not involve payment, but instead tap into the enthusiasm of participants to advance science. Typical tasks in these projects are small, well-defined, and carried out entirely online (e.g., classifying galaxies, identifying animals from digital photographs) (Garcia-Molina et al., 2016).

Commonly, micro-tasks are stripped of broader context and underlying objectives, and can be treated by workers as stand-alone autonomous problems (Brynjolfsson, Geva, & Reichman, 2016; Deng, Joshi, & Galliers, 2016). While special skills may be required to complete a task, broader understanding of the projects or sponsoring organizations is not expected. Each task accrues some cost – making it desirable to "minimize the number of questions" (Parameswaran, Sarma, Garcia-Molina, Polyzotis, & Widom, 2011, p. 267) and trade off cost against other objectives, such as accuracy, completeness, timeliness (see, e.g., Amsterdamer & Milo, 2015). The questions posted to the crowd may have widely-accepted answers that deal with general knowledge (e.g., "What is the capital of country X?") or be more subjective and individual (e.g., "Is movie X a comedy?") (see (Amsterdamer & Milo, 2015)).

As task-based crowdsourcing seeks to gather evidence for well-defined problems or tasks, it conforms to deductive reasoning and the hypothetico-deductive method of inquiry (Evans, Newstead, & Byrne, 1993; Popper, 2002). Deductive reasoning relies on making general statements (i.e., philosophically, beginning with a premise) and finding specific instances of these general statements. The hypothetico-deductive method proceeds by formulating a hypothesis in a form that could be falsified by empirical evidence. As a result, many researchers suggest that crowdsourcing (being a typical data collection endevour that involves people) should be targeted and hypothesis-driven (Francis, Blancher, & Phoenix, 2009; Goodman & Paolacci, 2017; Nichols & Williams, 2006; Wiersma, 2005).

Task-based crowdsourcing naturally fits the prevailing conceptualization of knowledge inquiry and data collection. It also aligns with prevailing methods in science, where conclusions derived primarily from exploratory analysis are still met with skeptisism under the typical assumption that "[i]t is a capital mistake to look at the data before you have identified all the theories" (Leamer, 1990, p. 239). Deductive reasoning is further aligned with traditional data management technologies and approaches (e.g., relational databases, entity relationship diagrams) based on a *closed world* assumption (Reiter, 1987).

While task-based crowdsourcing remains an important area, data management knowledge can be expanded through deeper understanding of another major model of crowdsourcing, observational crowdsourcing, in which organizations harness human perceptual and information-gathering abilities to make sense of the environment in which they operate. Characteristic features of this type of crowdsourcing include:

1. Long-term, continuous data collection;
2. Ill-defined, open-ended tasks;
3. At least in part performed out in the world;
4. Mostly voluntary in nature.

Major examples of observational crowdsourcing include eBird.org, iSpotNature.org, CitySourced.com, and MedWatch (www.fda.gov/Safety/MedWatch). Applications of these projects span diverse topics such as community mapping, crisis management, civic engagement, corporate market surveillance, and online citizen science. These projects are typically created by organizations as standalone platforms to allow people to report on natural or social phenomena they experience in the course of daily life, as well as to interact with each other and the organizational sponsors.

To illustrate the breadth of observational crowdsourcing, consider one of its major manifestations, citizen science, in which organizations (e.g., academic institutions) create purpose-built, open participation, online platforms that allow non-experts to provide data, view data, and often interact with scientists (Gura, 2013; Levy & Germonprez, 2017; Show, 2015; Wiersma, 2010). For example, in 2002 Cornell University launched eBird (www.ebird.com) to collect bird sightings from amateur birdwatchers around the globe to generate data for their ornithology research program (Hochachka et al., 2012; Sullivan et al., 2009).[1] Similar scientific projects include GalaxyZoo (galaxies), iSpotNature.org (global natural history with an historic UK focus), and many regional projects. To appreciate the pervasiveness of citizen science, consider that there are over a dozen large scale, global projects that focus on soils alone (e.g., OPAL Soil, The GLOBE, mySoil App, see (Rossiter, Liu, Carlisle, & Zhu, 2015)).

Much hope is vested in this kind of crowdsourcing, including assisting in tackling humanity's "evil quintet" – climate change, overexploitation, invasive species, land use change, and pollution (Theobald et al., 2015) – and unleashing unprecedented innovation and growth (Brynjolfsson & McAfee, 2014). In biology alone, it is estimated that more than two million people are engaged in major citizen science projects contributing up to $2.5 billion of in-kind value (Theobald et al., 2015). Citizen science promises to reduce research costs and has led to significant discoveries, such as Hanny's Voorwerp - a novel astronomical object named after a Dutch teacher Hanny van Arkel who discovered it on the GalaxyZoo project.

Indeed, citizen science was one of the first applications of crowdsourcing. Scientists have traditionally been accustomed to dealing with outside research participants, field settings, noisy data, and unreliable sources. Science has also been a major investor in resource-intensive and novel data collection infrastructure, and has embraced the promise of user-generated content and interactive web technologies since the late 1990s (Brossard, Lewenstein, & Bonney, 2005; Mims, 1999). Thus, the technological challenges (including those of data management) are relatively well established (Claire, Louise, & Mordechai, 2011; Wiggins et al., 2011).

Unlike highly focused, task-based crowdsourcing, which follows the hypothetico-deductive method of inquiry, observational crowdsourcing often relies on inductive reasoning. Philosophically, this makes (at least some) observational crowdsourcing data collection inductive in nature. Inductive reasoning involves hypothesis construction and making generalizations based on current empirical evidence, knowledge and understanding of phenomena. As a basis for systematic data collection, induction continues to be met with skepticism (Chalmers, 2013; Popper, 2002). Indeed, some observational crowdsourcing is targeted and hypothesis-based (Wiersma, 2010)

At the same time, working with diverse volunteers who are asked to experience a complex and unbounded real world opens an exciting opportunity to discover something new, find unanticipated uses of data, and deepen our understanding of the domain. Consequently, researchers increasingly suggest to approach this type of crowdsourcing in a hypothesis-free manner and explicitly seek to capture "unknown unknowns" (Lukyanenko, Parsons, & Wiersma, 2016; Wintle, Runge, & Bekessy, 2010). Among other things, this suggests embracing a more open view of domains. The reliance on targeted collection, however, limits the potential for knowledge discovery using crowds. Any predefined data collection inherently constrains the variety of data that can be collected in a study by fixing in advance the classes and attributes of interest and, consequently, determines both the classes and attributes to be excluded. While this results in highly consistent data, it limits creativity, discovery and innovation in crowds.

As many traditional data management approaches (e.g., based on a closed-world assumption) are more aligned with task-based crowdsourcing, we call on researchers to support observational crowdsourcing with innovative data management solutions.

The differences between typical task-based and observational projects are summarized in Table 1. Next, we discuss trends and opportunities in task-based crowdsourcing and then turn our attention to the observational crowdsourcing data management issues and challenges.

TASK-BASED CROWDSOURCING RESEARCH: TRENDS AND FUTURE OPPORTUNITIES

The work on task-based crowdsourcing highlights a number of problems requiring future research. Recently, there have been several reviews examining current and future potential of task-based crowdsourcing (Amsterdamer & Milo, 2015; Garcia-Molina et al., 2016; Li et al., 2016). Here, we briefly summarize some research questions in this stream (done here for context-setting and comparison with observational crowdsourcing, the main focus of this paper).

Major task-based crowdsourcing challenges relate to cost-benefit management when selecting the number of crowdworkers, number of tasks per worker, and models for aggregating crowd-produced results (Franklin et al., 2011). A major open challenge in crowdsourcing is incentivization and motivation of crowd workers (e.g., (Sheng et al., 2008)). As discussed by Amsterdamer and Milo (2015), trade-offs are inevitable and an unresolved issue is which objective function is best for a given application. Since incentives frequently involve payment, a research challenge is determining when and how much payment is appropriate for a given scenario, and when more intrinsic motivators can be more effective (Dow, Kulkarni, Klemmer, & Hartmann, 2012).

Information quality is a major challenge in task-based crowdsourcing. While crowd workers typically get paid for their contributions, the payments are often small, and traditional organizational incentives (e.g., awards, recognition by peers and promotions) are lacking (Deng et al., 2016; Gray, Suri, Ali, & Kulkarni, 2016). A common approach when dealing with small tasks assigned to crowd workers is that, by exploiting redundancy in the crowds (i.e., asking multiple users the same question, see (Franklin et al., 2011; Park & Widom, 2014)), some mean tendency can be established. Redundancy is thus a major form of quality control (in addition to other strategies, such as qualifying crowdworkers based on a number of previously completed tasks, and post-hoc statistical analysis, see (Hochachka et al., 2012; Wiggins et al., 2011)).

Another issue is crowd selection, as expertise and motivation differs substantially within crowds (Arazy et al., 2011; Coleman et al., 2009; Ogunseye & Parsons, 2016). As much task-based crowdsourcing is focusing on major platforms, such as Amazon's Mechanical Turk, research has examined crowd selection mechanisms provided by these platforms (e.g., qualifying tests, available worker profile variables) (Ipeirotis, Provost, & Wang, 2010; Kittur et al., 2008; Paolacci et al., 2010). A major stream of research examines task-worker matching and making task recommendations given worker attributes (e.g., expertise, past task performance) and declared preferences (Ambati, Vogel, &

Table 1. Comparison between typical task-based and observational crowdsourcing

Characteristic	Observational	Task-based
Nature of data collection	Long-term, continuous	Short, standalone
Problem formulation	Often ill-defined, open-ended	Well-defined
Data collection setting	Out in the world	Online
Nature of participation	Voluntary	Paid, sometimes voluntary
Philosophical underpinning	Inductive reasoning	Deductive reasoning

Carbonell, 2011; Geiger & Schader, 2014; Mao, Yang, Wang, Jia, & Harman, 2015; Yuen, King, & Leung, 2011, 2012). A promising, but thus far ill-explored (Amsterdamer et al., 2015), technique is specifying crowd properties directly when formulating a query. For example, a CrowdSQL (Franklin et al., 2011) query may be in this form:

```
SELECT SpeciesName
FROM BiologyPhotos
WHERE conservationStatus ~= 'Invasive'
   AND crowdExpertise IN ('Biology',  'Ecology', 'Australia')
```

In this query, the results are effectively conditioned on the right mix of crowd workers. This, of course, presupposes there is a shared ontology of crowd properties and leads to challenges in interpreting the results in the context of crowd qualifications.

Bringing humans into the technological chain of query formulation, writing, and execution necessitates rethinking existing query optimization algorithms. Specifically, query optimization strategies for crowds need to confront significantly greater latency and provide redundancy to deal with the unreliability of individual crowd worker results. The key to addressing this issue may lie in parallel execution and innovative approaches to decomposing, merging, and aggregating results (Difallah, Catasta, Demartini, & Cudré-Mauroux, 2014; Kucherbaev, Daniel, Tranquillini, & Marchese, 2016; Tranquillini, Daniel, Kucherbaev, & Casati, 2015).

A broader problem is establishing which tasks are amenable to human-powered intervention, discovering and enlarging the boundaries on crowd potential. An open issue is that of scale and the ability of human-powered problem-solving to handle big data queries. As the issue of information quality looms large when dealing with variable and often limited domain expertise, it motivates a search for innovative solutions for conceptual and logical storage and corresponding interfaces for micro-tasks that makes it easier for crowdworkers to find tasks to which they are able to contribute (Park & Widom, 2014).

We expect that research on task-based crowdsourcing will continue to benefit from major advances in related fields. Indeed, many studies of small tasks lie at the intersection of cognitive psychology, human-computer interaction, artificial intelligence (AI), natural language processing (NLP), and computer vision. A growing area of research is hybrid intelligence, wherein crowdsourcing is augmented with AI (Bernstein et al., 2015; Davis & Marcus, 2015; Guo et al., 2015). For example, a growing application of hybrid intelligence is when initial work is done by AI and humans are asked to tackle tasks for which AI has low confidence (e.g., beta version of CrowdFlowerAI, see crowdflower.com). As AI is less expensive than human crowds, a key challenge is in expanding abilities of AI to support human tasks. This can be done by using AI for pre-processing information and prioritizing work to be done by humans, and by identifying those tasks that are particularly difficult for AI, but easier for humans (Davis & Marcus, 2015; Guo et al., 2015).

Finally, research is needed to expand this context beyond such popular and well-understood platforms as Amazon's Mechanical Turk, and investigate micro-task problems in small projects where there is less redundancy and less is known about potential contributors (as an example, consider small projects created using EpiCollect.com[2]). In such scenarios, the "cold start" problem (i.e., when task requesters or the platforms cannot draw inferences for users about which it has not yet gathered sufficient information) is a major challenge (Lika, Kolomvatsos, & Hadjiefthymiades, 2014; Schein, Popescul, Ungar, & Pennock, 2002; Tejeda-Lorente, Bernabé-Moreno, Porcel, & Herrera-Viedma, 2017). Cold start potentially affects worker selection, providing appropriate interface designs, and selecting questions and conceptual structures matching potentially unknown characteristics of the crowd. These challenges motivate research examining potential solutions to this problem (Chen, Zhao, Wang, & Ng, 2016; Fan, Wei, Zhang, Yang, & Du, 2016; Safran & Che, 2017).

The challenges related to task-based crowdsourcing are summarized in Table 2, with sample solutions taken from research discussed above.

OBSERVATIONAL CROWDSOURCING: TRENDS AND OPPORTUNITIES

To understand data management challenges in observational crowdsourcing, consider the case of eBird. eBird contributors report birds they have seen in the wild. To generate consistent data, the collection is structured to make it amenable to search and retrieval by scientists. In particular, users are asked to select the biological species of the observed bird from a set of available options and indicate how many birds of the species were observed, as well as where and when the observation took place. Data are stored in a relational database.

Some decisions related to data management in a project such as eBird include: conceptual structures used to capture objects of interest to scientists; database schema used for storage; data collection interfaces; and, possibly, data visualization tools. Traditionally, users (in this case, amateur birders and scientists) would be involved in conceptual modeling and database design (for verification purposes) to ensure that the resulting structures reflect the views of all participants. The key issue is that, while scientists are a relatively homogenous and well understood stakeholder group, in a project such as eBird there are typically no constraints on who can participate. As a result, data contributors can be an extremely diverse, anonymous, and ill-understood user group. Therefore, it is unclear how to best structure data in such a project.

Managing Key Objective Trade-Offs: Quality, Utility and Participation

The eBird project shows a key challenge of observational crowdsourcing: managing the key trade-offs between utility of the data, data quality and user participation. In addition, recognizing that people may observe something unusual or novel, projects often have another objective: fostering discoveries.

It is generally recognized that data quality is a major issue for these types of projects (Kosmala et al., 2016; Lewandowski & Specht, 2015; Lukyanenko et al., 2014b). At the same time, eBird would not survive if volunteers do not provide observations of birds – making efforts to increase participation of great importance (Diner, Nakayama, Nov, & Porfiri, 2017; Domroese & Johnson, 2017; Lee, Crowston, Østerlund, & Miller, 2017). Finally, the project is designed to deliver data for scientific research and thus, citizen science projects organize domains to maximize the scientific utility of the data (Burgess et al., 2017; Prestopnik & Crowston, 2012; Stevens et al., 2014).

The search for solutions to address these three issues often means emphasizing one over the other. Thus, most projects in biology require positive identification (i.e., classification) of genera or species, as this classification level is useful for scientific analysis (Elbroch et al., 2011; Lewandowski & Specht,

Table 2. Challenges and sample solutions in task-based crowdsourcing

Challenge	Sample solutions
Task and worker cost-benefit management	Formal models of optimality between number of tasks per worker and task cost, models of optimal number of workers
Data quality management	Employing task redundancy, data post-processing, traning, tutorials
Incentives and motivation	Paying crowd workers, modeling optimal payment schedules
Crowd selection	Using CrowdSQL to specify crowd attributes, qualifying tests, platform-provided filters
Innovative interface design	Allowing workers to select questions to answer (e.g., CrowdFill)
Task-worker matching	Automatic task recommendation agents
Crowd SQL query optimization	Optimizing worker response time by decomposing tasks
Crowd-database integration	Human-powered queries (e.g., CrowdSQL)
Establishing bounds of crowdsourcing	Hybrid intelligence solutions

2015; Mayden, 2002). For example, eBird asks users to discriminate discriminate between Common, Caspian and Artic Tern. However, members of the general population targeted by crowdsourcing are not biology experts and may not be able to correctly identify at the species level (Gura, 2013; Lewandowski & Specht, 2015). Requiring contributors to classify at the species-genus level (or other levels of interest to science) may lead to guessing or project abandonment (Lukyanenko, Parsons, & Wiersma, 2014a; Lukyanenko et al., 2014b). One common solution to this challenge is traning data contributors, but this approach is limited given the often-anonymous Internet setting and uncommitted nature of online volunteers (Lewandowski & Specht, 2015). This is quite different from typical task-based crowdsourcing, where training may be required or crowd workers with desired characteristics can be specified as part of the task design.

Recently, researchers have been exploring advanced statistical techniques that can be used after data is captured to detect outliers or increase confidence in data provided by non-expert crowds (Bonney et al., 2014). Advanced computational techniques (e.g., live questions from an artificial agent) have also been used to support of volunteers during the data collection activity itself. For example, a promising direction is applying machine learning techniques that aim to classify an organism based on the photograph provided (e.g., eBird's MerlinApp). Yet, this is generally possible only with few target classes (which may not work in most general purpose projects[3]), and only when the photos are available and of sufficient resolution. In addition, such approaches do not recognize the value of capturing additional information about an observation. Thus, beyond classifying the focal object, projects are frequently interested in the context of observation and other notable objects. For example, a project may hope to collect in a structured form information such as "observed species *x* while on a tourist trip from *h* to *z*; *x* was preying on species *y*; species *y* appeared ill; one wing was broken and it seems to have been covered in oil; it was sunny and windy; species *x* and *y* were identified with the help of the tour guide *j*." As each such story would be different, this uniqueness makes observational crowdsourcing especially promising as a source of discoveries.

Data Sparsity and Data Intergration

From an organizational perspective, observational data production is continuous (consider eBird, which has been operating since 2002). At the same time, crowd members are free to drop in and out of projects sporadically, as organizations have little control over user participation (Rossiter et al., 2015; Wiersma, 2010). This results in a major challenge of integrating data across participants into a coherent domain view. Moreover, projects often involve observations of phenomena across a geographic area (e.g., UK in iSpot.org.uk, Australia in ala.org.au). This contrasts with the way much research has conceptualized crowdsourcing thus far, since, in task-based crowdsourcing, sparsity can be dealt with through task design. Indeed, the prevalence of redundancy as a common feature of task-based crowdsourcing creates an opposite challenge of data aggregation. Unlike micro-tasks that are necessarily focused, a major obstacle in observational data collection is integrating sparse and weakly-related data across time and space.

Large or Open Domain Boundaries

Traditional database development, as typically used in micro-task crowdsourcing, presupposes that domain boundaries are well-established. As organizations hope to better understand a complex and dynamic real-world environment through crowdsourcing, the scope of projects can be extensive and malleable. For example, iSpot.org.uk collected sightings of all-natural history in Great Britain. Similarly, GalaxyZoo images contain a variety of cosmic objects, some unknown to scientists themselves. A Canadian project, www.nlnature.com, interprets its interest in "local natural history" as including "any sighting of plants, animals, and other things (e.g., interesting rocks, landmarks)."[4] This means no single crowd member, developer or domain expert is likely to be an expert in the entire application domain.

Dealing with spatial, temporal, and attribute data scarcity is an important research direction. If an organization (e.g., a city) is looking to manage services based on a crowdsourcing data set (e.g., citysourced.com), it needs to "impute" gaps in coverage that necessarily arise when data is not collected systematically. The question arises whether it is possible to anticipate such gaps and proactively channel crowd effort in appropriate directions. Thus far, research has demonstrated the extent of this problem (Boakes et al., 2010; Lukyanenko et al., 2014b), but offered few solutions (none to our knowledge dealing with databases).

Need to Understand Data Creation Context

In many citizen science projects, it is critical to understand the context (i.e., relevant environmental factors) of data creation in order to better interpret the observations. In many projects, the phenomena about which users supply data may be available only to the original contributor (or a handful of people). For example, in projects that map biodiversity, the objects of interest (e.g., birds, animals) may be fleeting with a very short exposure time. In such cases, it is extremely difficult to exploit redundancy in the crowds (Franklin et al., 2011; Sheng et al., 2008), and the challenge is to get the most value out of a single data point. The anonymous nature of many projects further precludes seeking clarification or additional information unless such a query can be formulated before the contributor leaves the project.

While the issue is relevant to many projects, few specialized solutions exist. For example, a recent study suggests that the common practice of relying on the same volunteers over time may result in crowds becoming "stale"; this manifests in volunteers providing fewer context-rich observations (Ogunseye et al., 2017). This may suggest that projects should continuously recruit new volunteers, as well as offer incentives for existing volunteers to provide context-rich data.

Difficulty in Collecting System Requirements

A major feature of observational crowdsourcing is the democratic nature of participation. Projects such as eBird allow anyone to register and post observations of birds. While these projects are developed primarily to serve the needs of organizations (with stable cohorts of subject matter experts, such as scientists or product engineers), the users or contributors (i.e., citizen scientists, product consumers) are ordinary people, often lacking subject matter expertise and possessing diverse domain views (Gura, 2013; Lewandowski & Specht, 2015; Sullivan et al., 2009). Whereas task-based crowdsourcing may constrain who gets to perform a task (e.g., by eliminating members with low reputation, requiring a certain skill, or pruning data post-hoc based on known user attributes), this is often impossible when participation is open and anonymous. In citizen science, for example, many projects require only minimal information from participants (e.g., to comply with anonymity requirements of research protocols (see, e.g., Lukyanenko et al., 2014a)). As a result, some requirements and domain knowledge may originate from domain experts within organizations, but data are often provided by anonymous and non-expert users. It is furthermore impossible to reach every potential user who may wish to contribute data. As modelers are unable to fully determine the domain expertise of crowd providers and reach every potential user, a major challenge is how to conduct requirements elicitation and analysis to design appropriate database schemas and user interfaces that are congruent with the views of all users.

The challenges of requirements elicitation loom larger in observational crowdsourcing than in task-based projects. Unlike the latter, observational crowdsourcing projects are often developed from scratch and require custom modeling, database design and user interfaces. They are also considerably larger in development scope, and thus more expensive to implement, than task-based projects where existing platforms provide complete environments for worker participation and even provide ready-made templates for common tasks. In contrast, observational projects have to create the entire user environment for each project (e.g., data collection forms, visualization of existing data points, user forums). Furthermore, as many tasks are quite inexpensive, task providers working with such environments as AMT may freely experiment with test configurations until the desired outcomes are reached. In contrast, for many observational projects, there is a considerable delay (sometimes measured in years) between the original inception of the idea and the analysis of the data provided by contributors and the realization that something should be adjusted or modified in the task protocols and project designs. Consequently, it is critical to ensure that analysts collect accurate and complete requirements to specify appropriate features of the projects.

Motivated by the challenges to requirements elicitation in UGC, researchers have been exploring such solutions as user surrogates (e.g., usability experts), attracting average or representative users, conducing elicitation via videoconferencing platforms, and conducting requirements based on stratified sampling of users from common groups (e.g., non-experts vs. experts) (Anand & Mobasher, 2005; Carroll, Hoffman, Han, & Rosson, 2015; Iivari & Iivari, 2011; Iivari, 2011; Le Dantec, Asad, Misra, & Watkins, 2015; Salgado & Galanakis, 2014). Despite these solutions, often a very small group of users (e.g., those most accessible to the development team, such as scientists) are intensely involved (Bratteteig & Wagner, 2012, 2014). Requirements elicitation in UGC remains an open and exciting opportunity that researchers are continuously encouraged to pursue (Iivari, 2011; Kyng, 2010; Lukyanenko, Parsons, Wiersma, et al., 2016; Obendorf, Janneck, & Finck, 2009; Shapiro, 2010, 2010; Titlestad, Staring, & Braa, 2009).

Choice of Data Model

The choice of the data model affects the way information is collected and stored in crowdsourcing, including task difficulty and the quality of the resulting data. Thus, we specifically focus on the issue of data models in observational crowdsourcing.

Currently, there is no agreed-upon data model for observational crowdsourcing. Many major crowdsourcing projects, such as eBird, CitySourced, and EpiCollect, are based on traditional (e.g., relational) implementations, as there is no established solution specialized for observational crowdsourcing. Indeed, much work on micro-task based crowdsourcing extends the relational data model and SQL (Davidson, Khanna, Milo, & Roy, 2013; Franklin et al., 2011; Marcus, Wu, Karger, Madden, & Miller, 2011a). In contrast, some researchers argue the relational model has a negative impact on information quality and user participation and suggest that noSQL flexible data structures are more appropriate in this setting (Lukyanenko et al., 2014b).

In recent years, there has been growing interest in flexible data models. One approach is dynamic schema evolution or malleable schema (Dong & Halevy, 2005; Roussopoulos & Karagiannis, 2009). These approaches have been suggested for web-based systems and micro-task markets (Selke, Lofi, & Balke, 2012). As a variation, users may also be allowed to create their own schemas. The malleable schema approach, however, invites unresolved issues of cooperative schema evolution, concurrent access, and modification of schemas. It is also unclear if this approach is reliable for less knowledgeable or technology-averse users who may lack skills and motivation to create or alter models. Also, the sporadic nature of participation (anonymous or semi-anonymous in some projects) makes it difficult to assume quality of schemas (unlike in more predictable task-based crowdsourcing; see (Selke et al., 2012)).

One approach that has received research attention, including empirical evaluation, is the instance-based data model (Parsons & Wand, 2000) in crowdsourcing (Lukyanenko et al., 2017, 2014b). The model is based on the premise that instance and attribute are more fundamental modeling constructs than class. Although promising, unresolved issues with instance-based approaches include usability (i.e., developing appropriately flexible user interfaces) and assuring that data is useful for project sponsors. Given the demonstrable benefits of the instance-based model, an open question is whether other similar models, such as graph, key-value, or document-based, can also be leveraged (Cattell, 2011; Chang et al., 2008; DeCandia et al., 2007). Application of graph databases has been examined in task-based crowdsourcing (Parameswaran et al., 2011), but could be extended to observational crowdsourcing as well.

Both traditional and flexible approaches to modeling crowdsourced data embody the assumption that data schemas (rigid or flexible) are specified before data can be collected and stored. The case of citizen science above should motivate data management researchers to rethink this assumption.

One possible solution is storing user input in an unstructured form (e.g., as free-form text). However, without systematic data production in the direction required by project sponsors, the yield of useful information from free-form data collection is likely to be dismal. Furthermore, since a hallmark of crowdsourcing is massive datasets (even "big data") (Hochachka et al., 2012; Ipeirotis & Gabrilovich, 2014), to the extent possible data needs to be structured to allow for automatic querying and analysis. Currently many projects (e.g., eBird, GalaxyZoo) employ a hybrid solution: there are few predefined fields (e.g., type of galaxy; bird species observed), while additional information can be entered in a comments field, via email or an accompanying discussion forum (if available). This raises the question of how to determine the best allocation of data between structured and unstructured forms, and, importantly, how to increase the structured component of the project. Some solutions include post-hoc application of data-intensive techniques (e.g., data mining) (Hochachka et al., 2012). An open question is whether innovative data modeling and artificial intelligence can be used to collect more structured data at the point of capture. Artificial intelligence, including hybrid intelligence, is a major trend in task-based crowdsourcing (see above), yet its applications to observational projects remains limited.

Advances in natural language processing open opportunities in data collection and processing. One impact on data management research is increased utility of unstructured data formats when data from crowds can be collected as text and then parsed using text mining or natural language processing (NLP) (e.g., Jha, Andreas, Thadani, Rosenthal, & McKeown, 2010). A promising area of research resulting from NLP is voice-based data collection in crowdsourcing (e.g., dictating observations into a smart phone that is able to parse this information and possibly ask follow-up questions in real time). These applications may be limited in task-based environments (where activities occur primarily online), but may be quite effective in observational projects that involve generating data while experiencing the world. In many natural settings (in the forest, on the road, on a boat), speaking into a system may be more convenient than typing. Presently, we are not aware of research that specifically investigates this solution, creating an opportunity for a potential breakthrough in observational crowdsourcing.

Summary

Unlike task-based crowdsourcing that places crowds in a well-controlled technological chain, observational crowdsourcing conceptualizes crowds as integral elements of the organizational information chain that links internal decision making with the messy external environment. Crowds are sources of data and insights on the broader context in which organizations operate, giving people a more active and open-ended role in the data gathering process, and in the project as a whole. In this setting, database researchers are confronted with the need to produce solutions that are both technologically sound and effective at unlocking the potential of people as sensors of their surroundings, while delivering data of high quality to be usable in scientific research. The challenges related to observational crowdsourcing are summarized in Table 3; the table also presents potential solutions taken from research discussed above.

OUTLOOK FOR THE FUTURE

Crowdsourcing provides an opportunity for researchers to develop applications that support knowledge acquisition from diverse and heterogeneous crowds. Drawn to its promise, crowdsourcing research is on the rise. This survey of the crowdsourcing landscape suggests there are two major, but disconnected, "islands" of crowdsourcing research: task-based and observational. Task-based crowdsourcing builds innovative technological chains by inserting human workers at various points of data processing. In contrast, observational crowdsourcing conceptualizes as volunteers and "intelligent sensors" of reality.

Table 3. Challenges and solutions in observational crowdsourcing

Challenge	Sample solutions
Data quality	Training of volunteers, post-processing of data
Participation and motivation (as it relates to data management)	Gamification of tasks, removing participation barriers by making data collection easy and intuitive
Organizational (scientific) utility of data	Restricting data input, using advanced computational techniques to structure and analyze data after it is collected
Fostering discoveries	Promoting discoveries through open and flexible interfaces, asking real-time follow-up questions using natural language capable artificial agents
Optimizing objective trade-offs (e.g., between quality and utility)	Using open and flexible data collection processes and data models
Data sparsity, data intergration	Statistical techniques for data integration
Large and open domain boundaries	Democratizing requirements elicitation, keeping project open to new classes of things
Need to understand data creation context	Avoid crowd staleness, encourage volunteers to provide diverse and context-rich data
Difficulty in collecting system requirements	Leverage usability experts, representative users, massive videoconferencing platforms
What data model to use?	Flexible in schema-less solutions, NLP

While much scope for innovation remains in task-based crowdsourcing, research on data management for observational crowdsourcing is scarcer and the challenges formidable.

Our goal is to bring into focus the potential of observational crowdsourcing that we hope will catalyze database researchers to search for innovative solutions. While challenges in this domain are significant, we believe there are reasons to be optimistic. Many issues in observational crowdsourcing have been subjects of considerable research in the data management research community, including flexible schemas (Abiteboul, 1997; Angles & Gutierrez, 2008), data integration (Batini, Lenzerini, & Navathe, 1986; Zhao & Ram, 2007), schema evolution (Banerjee, Kim, Kim, & Korth, 1987; Liu, Chrysanthis, & Chang, 1994), data sparsity (Chang et al., 2008; Lehner, Albrecht, & Wedekind, 1998), conceptual modeling for dynamic and changing requirements (Chen, 2006; Liddle & Embley, 2007), data quality (Batini et al., 2015; Becker, 1998; Lee, Pipino, Strong, & Wang, 2004; Wang & Strong, 1996), and the rapidly expanding fields of AI, NLP, computer vision and hybrid intelligence (Bernstein et al., 2015; Davenport & Patil, 2012; Davis & Marcus, 2015; Nixon & Aguado, 2012). This and other relevant knowledge can be brought to bear on data management challenges in observational crowdsourcing.

The work in micro-task markets has already resulted in promising technologies that improve schema design, enhance query execution and optimization, and support innovative database and interface integration. Such innovative concepts as human-powered queries or joins have been proposed and implemented. Many of these solutions could be applied to power observational crowdsourcing as well. For example, the CrowdFill (Park & Widom, 2014) application introduces a novel data model for task-based crowdsourcing that gives workers considerable discretion in what to contribute. This data model could be leveraged in other kinds of crowdsourcing as well by giving volunteers flexibility in determining what kind of data they wish to contribute. Similarly, the concept of redundancy is central to many micro-tasks, with much ongoing research that examines efficient design of repeating tasks, organization of multiple labelers, detection of abnormalities, and models of optimality when other factors (e.g., cost) are considered (Franklin et al., 2011; Sheng et al., 2008). As many observational projects (e.g., eBird.org) actively employ statistical algorithms that seek to detect outliers and discover averages (Bonney et al., 2014; Hochachka et al., 2012), both streams of research stand to benefit from closer interaction.

There appears to be a significant opportunity for cross-pollination of findings and solutions between task-based and observational crowdsourcing. Both study very similar phenomena of large numbers of ordinary people contributing their abilities, talents and observational capacities to specific projects. This shared opportunity translates into a common challenge of creating an effective technological space that allows people of different views, motivations, and backgrounds to contribute successfully. Having this common challenge underscores the need for greater cooperation between the two islands of data management research in crowdsourcing.

The challenges posed by crowdsourcing (see Table 2 and Table 3) give rise to exciting research opportunities. The ongoing work on human-powered data management tasks has made significant strides, and many open questions remain. However, it is in another major type of crowdsourcing – one emphasizing human surveillance of the environment – that more wicked problems lurk. Unlike closed corporate environments which can be conceptually “frozen” to develop agreed-on conceptual and logical structures that represent domains, observational crowdsourcing is inherently dynamic and open. This translates into an opportunity for researchers to gain a deeper understanding of crowdsourcing and leverage this knowledge to potentially revolutionize data management theory and practice.

REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, *26*(3), 1–12. doi:10.1145/1361684.1361685

Abiteboul, S. (1997). Querying semi-structured data. *Presented at the 6th International Conference on Database Theory - ICDT '97*. Heidelberg, Germany: Springer.

Ambati, V., Vogel, S., & Carbonell, J. G. (2011). Towards task recommendation in micro-task markets. *Human Computation*, *11*.

Amsterdamer, Y., Davidson, S. B., Kukliansky, A., Milo, T., Novgorodov, S., & Somech, A. (2015). *Managing General and Individual Knowledge in Crowd Mining Applications*. CIDR.

Amsterdamer, Y., & Milo, T. (2015). Foundations of Crowd Data Sourcing. *SIGMOD Record*, *43*(4), 5–14. doi:10.1145/2737817.2737819

Anand, S., & Mobasher, B. (2005). Intelligent Techniques for Web Personalization. Springer. doi:10.1007/11577935_1

Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys*, *40*(1), 1–39. doi:10.1145/1322432.1322433

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *Journal of Management Information Systems*, *27*(4), 71–98. doi:10.2753/MIS0742-1222270403

Banerjee, J., Kim, W., Kim, H.-J., & Korth, H. F. (1987). *Semantics and implementation of schema evolution in object-oriented databases* (Vol. 16). ACM. doi:10.1145/38713.38748

Barwise, P., & Meehan, S. (2010). The One Thing You Must Get Right When Building a Brand. *Harvard Business Review*, *88*(12), 80–84.

Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *Computer Survey*, *18*(4), 323–364. doi:10.1145/27633.27634

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From Data Quality to Big Data Quality. [JDM]. *Journal of Database Management*, *26*(1), 60–82. doi:10.4018/JDM.2015010103

Becker, S. (1998). A practical perspective on data quality issues. *Journal of Database Management*, *9*(1), 35–37. doi:10.4018/jdm.1998010105

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., & Panovich, K. et al. (2015). Soylent: A Word Processor with a Crowd Inside. *Communications of the ACM*, *58*(8), 85–94. doi:10.1145/2791285

Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6), 1–11. doi:10.1371/journal.pbio.1000385 PMID:20532234

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science*, *343*(6178), 1436–1437. doi:10.1126/science.1251554 PMID:24675940

Brabham, D. C. (2013). *Crowdsourcing*. Cambridge, MA: MIT Press.

Bratteteig, T., & Wagner, I. (2012). Disentangling power and decision-making in participatory design (pp. 41–50). *Presented at the 12th Participatory Design Conference: Research Papers*. ACM.

Bratteteig, T., & Wagner, I. (2014). *Disentangling Participation: Power and Decision-making in Participatory Design*. New York, NY: Springer International Publishing. doi:10.1007/978-3-319-06163-4

Brossard, D., Lewenstein, B., & Bonney, R. (2005). Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education*, *27*(9), 1099–1121. doi:10.1080/09500690500069483

Brynjolfsson, E., Geva, T., & Reichman, S. (2016). Crowd-Squared: Amplifying the Predictive Power of Search Trend Data. *Management Information Systems Quarterly*, *40*(4), 941–961. doi:10.25300/MISQ/2016/40.4.07

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. New York, NY: WW Norton & Company.

Burgess, H., DeBey, L., Froehlich, H., Schmidt, N., Theobald, E., Ettinger, A., & Parrish, J. et al. (2017). The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C., & Nichol, R. C. et al. (2009). Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society*, *399*(3), 1191–1205. doi:10.1111/j.1365-2966.2009.15383.x

Carroll, J. M., Hoffman, B., Han, K., & Rosson, M. B. (2015). Reviving community networks: Hyperlocality and suprathresholding in Web 2.0 designs. *Personal and Ubiquitous Computing*, *19*(2), 477–491. doi:10.1007/s00779-014-0831-y

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *SIGMOD Record*, *39*(4), 12–27. doi:10.1145/1978915.1978919

Chalmers, A. F. (2013). *What is this thing called science?* Hackett Publishing.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., & Gruber, R. E. et al. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems*, *26*(2), 4–23. doi:10.1145/1365815.1365816

Chen, J., Xu, H., & Whinston, A. B. (2011). Moderated online communities and quality of user-generated content. *Journal of Management Information Systems*, *28*(2), 237–268. doi:10.2753/MIS0742-1222280209

Chen, P. (2006). Suggested Research Directions for a New Frontier – Active Conceptual Modeling. In ER 2006 (pp. 1–4).

Chen, W., Zhao, Z., Wang, X., & Ng, W. (2016). Crowdsourced Query Processing on Microblogs. *Presented at the International Conference on Database Systems for Advanced Applications* (pp. 18–32). Springer. doi:10.1007/978-3-319-32025-0_2

Claire, E., Louise, F., & Mordechai, H. (2011). A Flexible Database-Centric Platform for Citizen Science Data Capture. In *Proceedings of the IEEE Seventh International Conference on e-Science Workshops,* Stockholm, Sweden (pp. 39–44).

Clery, D. (2011). Galaxy Zoo volunteers share pain and glory of research. *Science*, *333*(6039), 173–175. doi:10.1126/science.333.6039.173 PMID:21737731

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, *4*(1), 332–358.

Daugherty, T., Eastin, M., & Bright, L. (2008). Exploring Consumer Motivations for Creating User-Generated Content. *Journal of Interactive Advertising*, *8*(2), 16–25. doi:10.1080/15252019.2008.10722139

Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review*, *90*, 70–76. PMID:23074866

Davidson, S. B., Khanna, S., Milo, T., & Roy, S. (2013). Using the crowd for top-k and group-by queries. *Presented at the 16th International Conference on Database Theory* (pp. 225–236). ACM. doi:10.1145/2448496.2448524

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103. doi:10.1145/2701413

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., & Vogels, W. et al. (2007). Dynamo: Amazon's highly available key-value store. *Operating Systems Review*, *41*(6), 205–220.

Delort, J.-Y., Arunasalam, B., & Paris, C. (2011). Automatic Moderation of Online Discussion Sites. *International Journal of Electronic Commerce*, *15*(3), 9–30. doi:10.2753/JEC1086-4415150302

Deng, X., Joshi, K., & Galliers, R. D. (2016). The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful through Value Sensitive Design. *Management Information Systems Quarterly*, *40*(2), 279–302. doi:10.25300/MISQ/2016/40.2.01

Difallah, D. E., Catasta, M., Demartini, G., & Cudré-Mauroux, P. (2014). Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. *Presented at the Second AAAI Conference on Human Computation and Crowdsourcing*.

Diner, D., Nakayama, S., Nov, O., & Porfiri, M. (2017). Social signals as design interventions for enhancing citizen science contributions. *Information Communication and Society*.

Dissanayake, I., Zhang, J., & Gu, B. (2015). Task division for team success in crowdsourcing contests: Resource allocation and alignment effects. *Journal of Management Information Systems*, *32*(2), 8–39. doi:10.1080/07421222.2015.1068604

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54*(4), 86–96. doi:10.1145/1924421.1924442

Domroese, M. C., & Johnson, E. A. (2017). Why watch bees? Motivations of citizen science volunteers in the Great Pollinator Project. *Biological Conservation*, *208*, 40–47. doi:10.1016/j.biocon.2016.08.020

Dong, X., & Halevy, A. (2005). *Malleable schemas: A preliminary report* (pp. 139–144). Proc. of WebDB.

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work (pp. 1013–1022). *Presented at the ACM 2012 conference on Computer Supported Cooperative Work*. ACM.

Elbroch, M., Mwampamba, T. H., Santos, M. J., Zylberberg, M., Liebenberg, L., Minye, J., & Reddy, E. et al. (2011). The Value, Limitations, and Challenges of Employing Local Experts in Conservation Research. *Conservation Biology*, *25*(6), 1195–1202. doi:10.1111/j.1523-1739.2011.01740.x PMID:21966985

Evans, J. S. B., Newstead, S. E., & Byrne, R. M. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Fan, J., Wei, Z., Zhang, D., Yang, J., & Du, X. (2016). Distribution-aware crowdsourced entity collection. *IEEE Transactions on Knowledge and Data Engineering*. doi:10.1109/TKDE.2016.2611509

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, *22*(5), 1224–1239. doi:10.1287/orsc.1100.0614

Francis, C. M., Blancher, P. J., & Phoenix, R. D. (2009). Bird monitoring programs in Ontario: What have we got and what do we need? *Forestry Chronicle*, *85*(2), 202–217. doi:10.5558/tfc85202-2

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011). CrowdDB: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 61–72). Athens, Greece: ACM. doi:10.1145/1989323.1989331

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in Data Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, *28*(4), 901-911. doi:10.1109/TKDE.2016.2518669

Geiger, D., & Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems—Current state of the art. *Decision Support Systems*, *65*, 3-16. doi:10.1016/j.dss.2014.05.007

Germonprez, M., & Hovorka, D. S. (2013). Member engagement within digitally enabled social network communities: New methodological considerations. *Information Systems Journal*, *23*(6), 525–549. doi:10.1111/isj.12021

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *The Journal of Consumer Research*, *44*(1), 196–210. doi:10.1093/jcr/ucx047

Gray, M. L., Suri, S., Ali, S. S., & Kulkarni, D. (2016). The crowd is a collaborative network (pp. 134–147). *Presented at the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.

Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N. Y., Huang, R., & Zhou, X. (2015). Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys*, *48*(1), 7. doi:10.1145/2794400

Gura, T. (2013). Citizen science: Amateur experts. *Nature*, *496*(7444), 259–261. doi:10.1038/nj7444-259a PMID:23586092

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), 130–137. doi:10.1016/j.tree.2011.11.006 PMID:22192976

Iivari, J., & Iivari, N. (2011). Varieties of user-centredness: An analysis of four systems development methods. *Information Systems Journal*, *21*(2), 125–153. doi:10.1111/j.1365-2575.2010.00351.x

Iivari, N. (2011). Participatory design in OSS development: Interpretive case studies in company and community OSS development contexts. *Behaviour & Information Technology*, *30*(3), 309–323. doi:10.1080/0144929X.2010.503351

Ipeirotis, P. G., & Gabrilovich, E. (2014). Quizz: Targeted crowdsourcing with a billion (potential) users. *Presented at the 23rd international conference on World wide web* (pp. 143–154). International World Wide Web Conferences Steering Committee. doi:10.1145/2566486.2567988

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 64–67). ACM. doi:10.1145/1837885.1837906

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94. doi:10.1145/2611567

Jha, M., Andreas, J., Thadani, K., Rosenthal, S., & McKeown, K. (2010). Corpus creation for new genres: A crowdsourced approach to PP attachment. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 13–20). Association for Computational Linguistics.

Kane, G. C., Alavi, M., Labianca, G. J., & Borgatti, S. (2014). What's different about social media networks? A framework and research agenda. *Management Information Systems Quarterly*, *38*(1), 274–304. doi:10.25300/MISQ/2014/38.1.13

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Presented at the SIGCHI conference on human factors in computing systems* (pp. 453–456). ACM.

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. doi:10.1002/fee.1436

Kucherbaev, P., Daniel, F., Tranquillini, S., & Marchese, M. (2016). ReLauncher: crowdsourcing micro-tasks runtime controller. *Presented at the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1609–1614). ACM.

Kyng, M. (2010). Bridging the Gap Between Politics and Techniques. *Scandinavian Journal of Information Systems*, *22*(1), 49–68.

Le Dantec, C. A., Asad, M., Misra, A., & Watkins, K. E. (2015). Planning with crowdsourced data: rhetoric and representation in transportation planning. *Presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1717–1727). ACM. doi:10.1145/2675133.2675212

Leamer, E. E. (1990). Specification problems in econometrics. In Econometrics: The New Palgrave (pp. 238–245). Springer. doi:10.1007/978-1-349-20570-7_32

Lee, T. K., Crowston, K., Østerlund, C. S., & Miller, G. (2017). Recruiting Messages Matter: Message Strategies to Attract Citizen Scientists. *Presented at the CSCW Companion* (pp. 227–230).

Lee, Y. W., Pipino, L., Strong, D. M., & Wang, R. Y. (2004). Process-embedded data integrity. *Journal of Database Management*, *15*(1), 87–103. doi:10.4018/jdm.2004010104

Lehner, W., Albrecht, J., & Wedekind, H. (1998). Normal forms for multidimensional databases. *Presented at the Tenth International Conference onScientific and Statistical Database Management* (pp. 63–72). IEEE.

Leimeister, J. M., Huber, M., Bretschneider, U., & Krcmar, H. (2009). Leveraging crowdsourcing: Activation-supporting components for IT-based ideas competition. *Journal of Management Information Systems*, *26*(1), 197–224. doi:10.2753/MIS0742-1222260108

Levina, N., & Arriaga, M. (2014). Distinction and Status Production on User-Generated Content Platforms: Using Bourdieu's Theory of Cultural Production to Understand Social Dynamics in Online Fields. *Information Systems Research*, *25*(3), 468–488. doi:10.1287/isre.2014.0535

Levy, M., & Germonprez, M. (2017). The Potential for Citizen Science in Information Systems Research. *Communications of the Association for Information Systems*, *40*(1), 2.

Lewandowski, E., & Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology*, *29*(3), 713–723. doi:10.1111/cobi.12481 PMID:25800171

Li, G., Wang, J., Zheng, Y., & Franklin, M. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Liddle, S. W., & Embley, D. W. (2007). A common core for active conceptual modeling for learning from surprises. In P. C. Peter & Y. W. Leah (Eds.), International Workshop on Active Conceputal Modeling of Learning (pp. 47–56). Springer-Verlag. doi:10.1007/978-3-540-77503-4_5

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073. doi:10.1016/j.eswa.2013.09.005

Liu, C.-T., Chrysanthis, P. K., & Chang, S.-K. (1994). Database schema evolution through the specification and maintenance of changes on entities and relationships. In P. Loucopoulos (Ed.), International Conference on Conceptual Modeling (pp. 132–151). Springer. doi:10.1007/3-540-58786-1_77

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2011). Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd. In H. Jain, A. Sinha, & P. Vitharana (Eds.), DESRIST (pp. 465–473). Springer Berlin / Heidelberg.

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014a). The Impact of Conceptual Modeling on Dataset Completeness: A Field Experiment. In *Proceedings of the International Conference on Information Systems (ICIS)*.

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014b). The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-generated Content. *Information Systems Research*, *25*(4), 669–689. doi:10.1287/isre.2014.0537

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2016). Emerging problems of data quality in citizen science. *Conservation Biology*, *30*(3), 447–449. doi:10.1111/cobi.12706 PMID:26892841

Lukyanenko, R., Parsons, J., Wiersma, Y., Sieber, R., & Maddah, M. (2016). Participatory Design for User-generated Content: Understanding the challenges and moving forward. *Scandinavian Journal of Information Systems*, *28*(1), 37–70.

Lukyanenko, R., Parsons, J., Wiersma, Y. F., Wachinger, G., Huber, B., & Meldt, R. (2017). Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-generated Content. *Journal of the Association for Information Systems*, *18*(4), 297–339.

Majchrzak, A. N. N., & More, P. H. B. (2011). Emergency! Web 2.0 to the Rescue! *Communications of the ACM*, *54*(4), 125–132. doi:10.1145/1924421.1924449

Mao, K., Yang, Y., Wang, Q., Jia, Y., & Harman, M. (2015). Developer recommendation for crowdsourced software development tasks. *Presented at the 2015 IEEE Symposium on Service-Oriented System Engineering (SOSE)* (pp. 347–356). IEEE.

Marcus, A., Wu, E., Karger, D. R., Madden, S., & Miller, R. C. (2011a). *Crowdsourced databases: Query processing with people* (pp. 211–214). CIDR.

Marcus, A., Wu, E., Karger, D. R., Madden, S., & Miller, R. C. (2011b). Human-powered sorts and joins. *VLDB Endowment*, *5*(1), 13–24. doi:10.14778/2047485.2047487

Mayden, R. L. (2002). On biological species, species concepts and individuation in the natural world. *Fish and Fisheries*, *3*(3), 171–196. doi:10.1046/j.1467-2979.2002.00086.x

Mims, F. M. III. (1999). Amateur science - Strong tradition, bright future. *Science*, *284*(5411), 55–56. doi:10.1126/science.284.5411.55

Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in Ecology & Evolution*, *21*(12), 668–673. doi:10.1016/j.tree.2006.08.007 PMID:16919361

Nixon, M. S., & Aguado, A. S. (2012). *Feature Extraction & Image Processing for Computer Vision*. Waltham, MA: Academic Press.

Nov, O., Arazy, O., & Anderson, D. (2011). Dusting for science: motivation and participation of digital citizen science volunteers. In *2011 IConference* (pp. 68-74).

Obendorf, H., Janneck, M., & Finck, M. (2009). Inter-contextual distributed participatory design. *Scandinavian Journal of Information Systems*, *21*(1), 2.

Ogunseye, S., & Parsons, J. (2016). Can Expertise Impair the Quality of Crowdsourced Data? In SIGOPEN Developmental Workshop at ICIS 2016.

Ogunseye, S., Parsons, J., & Lukyanenko, R. (2017). Do Crowds Go Stale? Exploring the Effects of Crowd Reuse on Data Diversity. In WITS 2017, Seoul, South Korea.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Parameswaran, A., Sarma, A. D., Garcia-Molina, H., Polyzotis, N., & Widom, J. (2011). Human-assisted graph search: It’s okay to ask questions. *VLDB Endowment*, *4*(5), 267–278. doi:10.14778/1952376.1952377

Parameswaran, A. G., Park, H., Garcia-Molina, H., Polyzotis, N., & Widom, J. (2012). Deco: declarative crowdsourcing. *Presented at the 21st ACM international conference on Information and knowledge management* (pp. 1203–1212). ACM.

Park, H., Pang, R., Parameswaran, A., Garcia-Molina, H., Polyzotis, N., & Widom, J. (2013). An overview of the deco system: Data model and query language; query processing and optimization. *SIGMOD Record*, *41*(4), 22–27. doi:10.1145/2430456.2430462

Park, H., & Widom, J. (2014). Crowdfill: Collecting structured data from the crowd. In *ACM SIGMOD International Conference on Management of Data* (pp. 577–588).

Parsons, J., Lukyanenko, R., & Wiersma, Y. (2011). Easier citizen science is better. *Nature*, *471*(7336), 37–37. doi:10.1038/471037a PMID:21368811

Parsons, J., & Wand, Y. (2000). Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems*, *25*(2), 228–268. doi:10.1145/357775.357778

Popper, K. (2002). *The Logic of Scientific Discovery*. London, UK: Taylor & Francis; Retrieved from http://books.google.ca/books?id=T76Zd20IYlgC

Prestopnik, N., & Crowston, K. (2011). Gaming for (Citizen) Science: Exploring Motivation and Data Quality in the Context of Crowdsourced Science through the Design and Evaluation of a Social-Computational System. In *Proceedings of the IEEE International Conference on e-Science Workshops (eScienceW)* (pp. 1–28). doi:10.1109/eScienceW.2011.14

Prestopnik, N., & Crowston, K. (2012). Purposeful gaming & socio-computational systems: a citizen science design case. *Presented at the 17th ACM international conference on Supporting group work* (pp. 75–84). ACM. doi:10.1145/2389176.2389188

Reiter, R. (1987). On closed world data bases. In M. L. Ginsberg (Ed.), Readings in nonmonotonic reasoning (pp. 300–310). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=42641.42663

Rossiter, D. G., Liu, J., Carlisle, S., & Zhu, A.-X. (2015). Can citizen science assist digital soil mapping? *Geoderma*, *259*, 71–80. doi:10.1016/j.geoderma.2015.05.006

Roussopoulos, N., & Karagiannis, D. (2009). Conceptual Modeling: Past, Present and the Continuum of the Future. In A. Borgida, C.V.P. Giorgini, & E. Yu (Eds.), Conceptual Modeling: Foundations and Applications (pp. 139–152). Berlin / Heidelberg: Springer.

Safran, M., & Che, D. (2017). Real-time recommendation algorithms for crowdsourcing systems. *Applied Computing and Informatics*, *13*(1), 47–56. doi:10.1016/j.aci.2016.01.001

Salgado, M., & Galanakis, M. (2014). ... so what?: limitations of participatory design on decision-making in urban planning. *Presented at the Participatory Design Conference* (pp. 5–8). ACM. doi:10.1145/2662155.2662177

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *Presented at the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 253–260). ACM.

Selke, J., Lofi, C., & Balke, W.-T. (2012). Pushing the boundaries of crowd-enabled databases with query-driven schema expansion. *VLDB Endowment*, *5*(6), 538–549. doi:10.14778/2168651.2168655

Shapiro, D. (2010). A Modernised Participatory Design? A reply to Kyng. *Scandinavian Journal of Information Systems*, *22*(1), 69–76.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In ACM SIG KDD (pp. 614–622).

Sheppard, S., Wiggins, A., & Terveen, L. (2014). Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data. *Presented at the ACM conference on Computer supported cooperative work & social computing* (pp. 1234–1245).

Show, H. (2015). Rise of the citizen scientist. *Nature*, *524*(7565), 265–265. doi:10.1038/524265a PMID:26289171

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. *Presented at the companion publication of the 23rd international conference on World wide web companion* (pp. 1049–1054). International World Wide Web Conferences Steering Committee. doi:10.1145/2567948.2579215

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops CVPRW '08*. doi:10.1109/CVPRW.2008.4562953

Stevens, M., Vitos, M., Altenbuchner, J., Conquest, G., Lewis, J., & Haklay, M. (2014). Taking participatory citizen science to extremes. *Pervasive Computing, IEEE*, *13*(2), 20–29. doi:10.1109/MPRV.2014.37

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. doi:10.1016/j.biocon.2009.05.006

Tejeda-Lorente, A., Bernabé-Moreno, J., Porcel, C., & Herrera-Viedma, E. (2017). Using Bibliometrics and Fuzzy Linguistic Modeling to Deal with Cold Start in Recommender Systems for Digital Libraries. In *Advances in Fuzzy Logic and Technology 2017* (pp. 393–404). Springer.

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., & Harsch, M. A. et al. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. doi:10.1016/j.biocon.2014.10.021

Titlestad, O. H., Staring, K., & Braa, J. (2009). Distributed development to enable user participation: Multilevel design in the HISP network. *Scandinavian Journal of Information Systems*, *21*(1), 3.

Tranquillini, S., Daniel, F., Kucherbaev, P., & Casati, F. (2015). BPMN task instance streaming for efficient micro-task crowdsourcing processes (pp. 333–349). *Presented at the International Conference on Business Process Management*. Springer. doi:10.1007/978-3-319-23063-4_23

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. doi:10.1080/07421222.1996.11518099

Weld, D. S. (2015). Intelligent Control of Crowdsourcing. *Presented at the 20th International Conference on Intelligent User Interfaces*. ACM.

Wiersma, Y. F. (2005). Environmental benchmarks vs. ecological benchmarks for assessment and monitoring in Canada: Is there a difference? *Environmental Monitoring and Assessment*, *100*(1–3). doi:10.1007/s10661-005-7055-6 PMID:15727295

Wiersma, Y. F. (2010). Birding 2.0: Citizen science and effective monitoring in the Web 2.0 world. *Avian Conservation & Ecology*, *5*(2), 13. doi:10.5751/ACE-00427-050213

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., & LeBuhn, G. … Newman, G. (2013). Data management guide for public participation in scientific research.

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science. In "Computing for Citizen Science" workshop (pp. 14–19). Stockholm, SE. doi:10.1109/eScienceW.2011.27

Wintle, B. A., Runge, M. C., & Bekessy, S. A. (2010). Allocating monitoring effort in the face of unknown unknowns. *Ecology Letters*, *13*(11), 1325–1337. doi:10.1111/j.1461-0248.2010.01514.x PMID:20678146

Yuen, M.-C., King, I., & Leung, K.-S. (2011). Task matching in crowdsourcing. *Presented at the 2011 International Conference on Internet of Things (iThings/CPSCom) and 4th International Conference on Cyber, Physical and Social Computing* (pp. 409–412). IEEE.

Yuen, M.-C., King, I., & Leung, K.-S. (2012). Task recommendation in crowdsourcing systems. *Presented at the first international workshop on crowdsourcing and data mining* (pp. 22–26). ACM. doi:10.1145/2442657.2442661

Zhao, H., & Ram, S. (2007). Combining schema and instance information for integrating heterogeneous data sources. *Data & Knowledge Engineering*, *61*(2), 281–303. doi:10.1016/j.datak.2006.06.004

Zwass, V. (2010). Co-Creation: Toward a Taxonomy and an Integrated Research Perspective. *International Journal of Electronic Commerce*, *15*(1), 11–48. doi:10.2753/JEC1086-4415150101

ENDNOTES

[1] As of October 2017, eBird participants reported more than 400 million observations from around the world (Source: http://ebird.org/content/ebird/news/).

[2] See:http://www.isgtw.org/feature/collecting-data-gets-easier-epicollect.

[3] Currently Merlin App allows to identify 400 species (out of over 10,000 bird species), see http://merlin.allaboutbirds.org/help-and-faqs.

[4] http://www.nlnature.com/Endangered-Species-Biodiversity/About-NL-Nature-2.aspx.

Roman Lukyanenko is Assistant Professor of Information Systems in the Edwards School of Business, University of Saskatchewan. He received his PhD in information systems from Memorial University of Newfoundland. His research interests include citizen science, crowdsourcing, information (data) quality, conceptual modeling, and business intelligence. His research has been published in Nature, Information Systems Research, Journal of the Association for Information Systems, among others, as well as leading conferences in information systems and computer science.

Jeffrey Parsons is University Research Professor and Professor of Information Systems in the Faculty of Business Administration at Memorial University of Newfoundland. His research interests include information integration, crowdsourcing, information quality and conceptual modeling. His research has been published broadly, including in journals such as Nature, Communications of the ACM, ACM Transactions on Database Systems, IEEE Transactions on Software Engineering and MIS Quarterly. He is a Senior Editor for MIS Quarterly. He has also served as Program Co-chair for a number of conferences, including AMCIS, WITS, and the ER conference. He holds a PhD in Information Systems from the University of British Columbia.