AMCIS 2020 Proceedings

Virtual Communities and Collaboration (VCC)

Aug 10th, 12:00 AM

What the Crowd Sources: A Protocol for a Contribution-Centred Systematic Literature Review of Data Crowdsourcing Research

Ryan J. A. Murphy
Memorial University of Newfoundland, rmurphy@mun.ca

Jeffrey Parsons
Memorial University of Newfoundland, jeffreyp@mun.ca

Follow this and additional works at: https://aisel.aisnet.org/amcis2020

What the Crowd Sources: A Protocol for a Contribution-Centered Systematic Literature Review of Data Crowdsourcing Research

Emergent Research Forum (ERF)

Ryan J. A. Murphy
Memorial University of Newfoundland
rmurphy@mun.ca

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Abstract

Data crowdsourcing is the mobilization of large groups of contributors—often volunteers via the Internet—to collect and/or analyze data. Research on data crowdsourcing often prioritizes the data consumer or project sponsor. Significant gaps remain in understanding how to address design issues from the perspective of data crowdsourcing contributors. A systematic literature review is an ideal method for identifying gaps in how researchers conceptualize contributions in data crowdsourcing. This project presents a protocol for such a systematic literature review of data crowdsourcing. We will use the protocol to guide a subsequent systematic literature review and the construction of a data-information-knowledge-wisdom chart that identifies critical gaps and opportunities for research in data crowdsourcing systems.

Keywords

Data crowdsourcing, systematic-literature review, contributors, user-generated content.

Crowdsourcing and data crowdsourcing

This paper describes the development of a protocol for a systematic literature review (SLR) of contribution-centred research on *data crowdsourcing*—a specific type of crowdsourcing in which contributors collect or analyze data for a coordinated purpose (Lukyanenko et al., 2017). While extensive research has been completed on data crowdsourcing, and many SLRs exist covering a variety of aspects of the field, no such SLR has provided a comprehensive overview of research focused on the contributor. An SLR will help to (a) systematically examine existing research on designing for contributors in data crowdsourcing, and (b) to establish a road map for future research directions, identifying important unanswered questions. Before an SLR can be conducted, however, an explicit, refined protocol for the review must be established. It is the objective of the present study to propose such a protocol.

Crowdsourcing has emerged as a method of mobilizing large numbers of people to accomplish a coordinated goal (Lukyanenko & Parsons, 2018). Data crowdsourcing is a crowdsourcing approach focused on data collection (Murphy & Parsons, 2020) and has become useful in many domains. For example, "citizen scientists" have helped advance HIV research through predicting novel ways to fold proteins (Cooper et al., 2010). A key concern for data crowdsourcing systems is maximizing the quality of information generated by the crowd (Lukyanenko, Parsons, & Wiersma, 2014). Our aim is to provide a systematic literature review of research centered on the data contributor and their contributions to data crowdsourcing.

At the core of crowdsourcing projects is a dialectic between the assigners (and subsequent users) of a crowd task and the data contributors to that task (Zhao & Zhu, 2014). Successful projects depend on the effective participation of these two actor groups. In data crowdsourcing, these respective roles take the forms of data targeting and consumption (typically by the project coordinator) and data contribution (via the crowd of contributors; (Lukyanenko et al., 2017). Given data consumers' leadership role in crowdsourcing projects, it may be sensible to evaluate the quality of a data crowdsourcing project from their perspective (e.g., Wang and Strong's conceptual framework of data quality; Wang & Strong, 1996). However, conceptualizations of

data quality from the perspective of the data contributor can also be valuable. Lukyanenko, Parsons, and Wiersma (2014), for example, found that a crowdsourcing platform that provided instance-and-attribute based inputs outperformed a class-based data collection system. They argued that the instance-based system provided an interface more suited to users of varying expertise, limiting data loss that would occur if users were uncertain of how to accurately classify their contributions in a class-based system.

Contributor-focused research on crowdsourcing

A variety of scholars have investigated the crowdsourcing data contributor and their contributions. Stewart (2010) differentiated crowdsourcing participants into three groups: super contributors, contributors, and outliers. A rich literature exists on contributor skill and motivation (Chandler & Kapelner, 2013; Kaufmann, Schulze, & Veit, 2011; Nov, Arazy, & Anderson, 2014; Zheng, Li, & Hou, 2011). In general, these authors find that crowd expertise and motivation are key factors influencing crowd participation. Other research directions explore how to design effective conceptual models to account for contributor heterogeneity (Lukyanenko et al., 2017; McGinnes, 2011). These designs may be useful in increasing flexibility for contributors without sacrificing the usability of data for consumers, particularly when combined with techniques such as machine learning (Lukyanenko, Parsons, Wiersma, & Maddah, 2019). Some work has attempted to synthesize what research has demonstrated about crowdsourcing contributors and their contributions in the form of typologies of crowdsourcing projects. Geiger, Rosemann, Fielt, and Schader (2012) analyzed crowdsourcing projects and found that contribution value differentiated along two axes: the diversity of forms of contributions (e.g., heterogeneous vs. homogeneous) and how value was derived from contributions (emergent vs. non-emergent). Similarly, Lukyanenko and Parsons (2018) provide a typology of data crowdsourcing projects that proposes two categorizations of what the crowd does (micro-task vs. observational crowdsourcing). Murphy and Parsons (2020) combine these two analyses to offer a data crowdsourcing system typology that emphasizes the contributor and their contributions, differentiating projects based on five design dimensions.

Ergo, designing for contributors has become a research stream in data crowdsourcing. However, while there have been many reviews of crowdsourcing research (e.g., Amrollahi & Ahmadi, 2019; Buettner, 2015; Correia, Schneider, Fonseca, & Paredes, 2018; Hossain & Kauranen, 2015; Mäkipää, Dang, Mäenpää, & Pasanen, 2020; Thuan, Antunes, & Johnstone, 2016), scholars have yet to produce a systematic review of research specifically on designing for contributors in data crowdsourcing, leaving a number of important questions unanswered. What are the best practices in designing for contributions in data crowdsourcing? What do we not know about contributors and contribution modelling in data crowdsourcing? Are there generally accepted overarching models or frameworks that govern design for contributors? If so, what about those models or frameworks remain to be established? In this research project, we aim to answer these research questions via a comprehensive analysis of the literature.

Towards a comprehensive review

A SLR is a stand-alone literature review that uses a systematic, rigorous standard (Okoli & Schabram, 2010). Unlike conventional literature reviews or theoretical backgrounds, SLRs go beyond providing a cursory overview of the field by synthesizing available material and critiquing theory. SLRs must follow a reproducible, explicit, and comprehensive method to find and assess existing research on a subject while synthesizing that research into a contribution valuable to the field (Okoli & Schabram, 2010).

In this paper, we present our protocol for this SLR by describing our SLR process according to Okoli and Schabram's (2010, p. 7) eight steps. Okoli and Schabram (2010, p. 35–36) suggest that publishing a review protocol in conference proceedings is an ideal step in conducting the review process. It provides an opportunity for both gathering feedback from reviewers and participants to refine the protocol and to validate the protocol externally before pursuing the research.

The purpose of the review

As per the research questions we posed above, we aim to (a) analyze the progress of research on designing for contributors in data crowdsourcing and (b) to establish a road map for future research directions, identifying important unanswered questions. A key contribution of the review will be the creation of a Data-

Information-Knowledge-Wisdom (DIKW) scheme (Ackoff, 1989; Sanders & Stappers, 2014, pp. 200-205) for design for contribution in data crowdsourcing. A DIKW scheme describes the formation of theory as a layered structure from the phenomena of the domain to the theory that governs predictions in that domain (Figure 1). We will identify the data, information, knowledge, and wisdom generated by scholars in design for contribution in data crowdsourcing by analyzing and coding literature according to the categories of the schema, and by visually linking the codes together in the DIKW scheme structure. The DIKW scheme will allow us to identify significant research sub-streams and gaps in the literature.
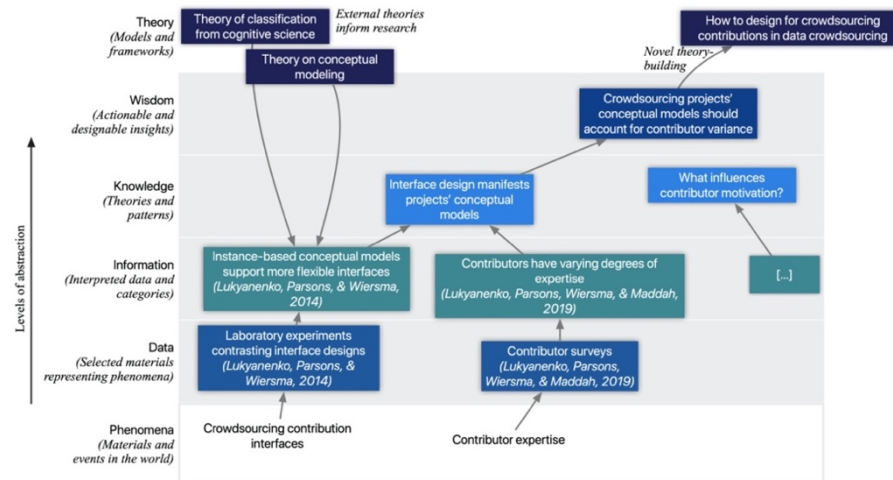

Figure 1. A DIKW scheme with sample concepts.

The literature search

To ensure we collect all papers relevant to contributor-focused data crowdsourcing research, we plan to conduct a broad literature search and use extensive inclusion/exclusion criteria to narrow paper selection.

Our search terms will include "crowdsourcing," "data crowdsourcing," "citizen science," and "volunteered geographic information" (as volunteered geographic information is a form of crowdsourcing; see e.g. (Horita, Degrossi, Assis, Zipf, & Albuquerque, 2013). We will use these search tokens to search the Association for Information Systems' (AIS) Basket of Journals (https://aisnet.org/page/SeniorScholarBasket), a select list of eight of the top journals in the field of Information Systems according to the College of Senior Scholars, as well as to search the AIS Special Interest Group Recommended Journals (*Ibid.*). We will also search the repositories of the major IS conferences: the Americas Conference on Information Systems (AMCIS), the European Conference on Information Systems (ECIS), the Hawai'i International Conference on Systems Sciences (HICSS), the International Conference on Information Systems (ICIS), and the Pacific Asia Conference on Information Systems (PACIS). After reviewers screen the resulting papers for inclusion and exclusion, we will use the papers' reference lists to conduct a one-level backwards search by examining the paper titles in these lists to identify any papers that were used by the semi-final papers not yet included. These papers will also be screened for inclusion and exclusion. Finally, considering that an SLR is a time-consuming process, we will repeat the search above as we near the conclusion of the review to collect any publications released between the completion of our initial search and submitting our review for publication.

Papers found via the search process will be downloaded to a database and reference management software (DEVONthink 3 Pro with a Bookends integration) with attached metadata to log the paper's progress through the subsequent steps. These include the paper's current review status (searched for, included, excluded, reviewed, data extracted), exclusion criteria (e.g., failed inclusion criteria, failed quality assessment), and exclusion rationale (e.g., "did not discuss anything related to contributions").

Practical screening (screening for inclusion)

We expect the search described above to return a broad collection of papers that mention crowdsourcing and the related topics included in our search criteria, but do not concern the topics of data crowdsourcing or design for contribution that are our focus. Thus, we will complete a scan of these papers to screen them for inclusion in the SLR. Reviewers will manually review the title and the abstract of each paper returned by our search to complete the practical screen. In addition to including only English-language papers, we will include any papers that have a focus on factors relating to the contributor (including motivation and expertise), a focus on data crowdsourcing (as opposed to e.g., crowdfunding platforms), and a focus on factors relating to contributions (e.g., those found in Murphy & Parsons' (2020) typology).

Quality screening (screening for exclusion)

Once reviewers complete the practical screening, papers will be evaluated based on their quality. The purpose we outlined for this SLR is understanding the state of research and identifying significant gaps in the literature. As such, assessing the logical arguments of the works will be of prime concern in our quality screening. We will appraise the quality of each paper by seeking out fallacies in what claims are made, what evidence supports them, and whether the evidence supports the claims (Okoli & Schabram, 2010).

Data extraction

Once reviewers finalize a collection of papers, we will examine the resulting papers for their contribution to the state of the literature. Reviewers will focus on the abstract, introduction, results, and discussion sections of each paper and take notes according to the following guidelines: (1) What are the research questions? (2) What are the assumptions about data crowdsourcing and design for contributors? (3) What do the authors predict about data crowdsourcing and design for contributors? (4) What do the authors find or conclude about data crowdsourcing and design for contributors? (5) What are the author's future research directions?

Data synthesis

The reviewers will iteratively qualitatively code the notes from data extraction using grounded field theory (Charmaz, 2006), with a focus on identifying how each paper relates to the broader literature. Here we adopt a "synthesis by explanation" approach to emphasize the causal linkages surfaced by the literature for design for contributors in data crowdsourcing (Okoli & Schabram, 2010, p. 32).

Discussion

This paper establishes a protocol for completing an SLR on designing for contributions in data crowdsourcing. As of this writing, according to Google Scholar, the first papers to mention data crowdsourcing began appearing just over a decade ago. As the field begins to mature, it makes sense to take stock of what questions researchers have asked and answered—and what have not. We focus on research in data crowdsourcing that emphasizes the role of the contributor and their contributions. We will examine the causal structure of published research in the field to construct a DIKW scheme. that will visualize what studies in this field have emphasized and what has been ignored. By mapping the gaps in research on this topic, we hope to guide future researchers to examine key issues that limit progress in data crowdsourcing.

REFERENCES

Ackoff, R. L. 1989. "From Data to Wisdom," *Journal of Applied Systems Analysis,* (16:1), pp. 3-9.

Amrollahi, A., & Ahmadi, M. H. 2019. "What Motivates the Crowd," *Crowdsourcing and Knowledge Management in Contemporary Business Environments: Advances in Logistics, Operations, and Management Science*, IGI Global, pp. 103-133. (doi:10.4018/978-1-5225-4200-1.ch006)

Buettner, R. 2015. "A Systematic Literature Review of Crowdsourcing Research From a Human Resource Management Perspective," in *Proceedings of the 48th Hawaii International Conference on System Sciences*, Kauai, HI, USA, p. 4609-4618. (doi:10.1109/hicss.2015.549)

Chandler, D., & Kapelner, A. 2013. "Breaking Monotony With Meaning: Motivation in Crowdsourcing Markets," *Journal of Economic Behavior & Organization,* (90), pp. 123-133. (doi:10.1016/j.jebo.2013.03.003)

Charmaz, K. 2006. *Constructing Grounded Theory,* London, UK: Sage Publications.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., & Players, F. 2010. "Predicting Protein Structures With a Multiplayer Online Game.," *Nature,* (466:7307), pp. 756-760. (doi:10.1038/nature09304)

Correia, A., Schneider, D., Fonseca, B., & Paredes, H. 2018. "Crowdsourcing and Massively Collaborative Science: A Systematic Literature Review and Mapping Study," in *Proceedings of the 24th International Conference on Collaboration and Technology*, A. Rodrigues, B. Fonseca, & N. Preguiça (eds.), Cham, Switzerland, p. 133-154. (doi:10.1007/978-3-319-99504-5_11)

Geiger, D., Rosemann, M., Fielt, E., & Schader, M. 2012. "Crowdsourcing Information Systems: Definition, Typology, and Design," in *Proceedings of the 33rd International Conference on Information Systems*, Orlando, USA, p. 1-11.

Horita, F. E., Degrossi, L., Assis, L. F. D., Zipf, A., & Albuquerque, J. P. D. 2013. "The Use of Volunteered Geographic Information (Vgi) and Crowdsourcing in Disaster Management: A Systematic Literature Review," *AMCIS 2013 Proceedings.*

Hossain, M., & Kauranen, I. 2015. "Crowdsourcing: A Systematic Literature Review," *Strategic Outsourcing.*

Kaufmann, N., Schulze, T., & Veit, D. 2011. "More Than Fun and Money. Worker Motivation in Crowdsourcing: A Study on Mechanical Turk," in *AMCIS 2011 Proceedings - All Submissions*, Detroit, Michigan, USA.

Lukyanenko, R., Parsons, J., Wiersma, Y., & Maddah, M. 2019. "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *Management Information Systems Quarterly,* (43:2), pp. 634-647.

Lukyanenko, R., & Parsons, J. 2018. "Beyond Micro-Tasks: Research Opportunities in Observational Crowdsourcing," *Journal of Database Management,* (29:1), pp. 1-22.

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

Lukyanenko, R., Wiersma, Y., Huber, B., Parsons, J., Wachinger, G., & Meldt, R. 2017. "Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-Generated Content," *Journal of the Association for Information Systems,* (18:4), pp. 297-339. (doi:10.17705/1jais.00456)

Mäkipää, J.-P., Dang, D., Mäenpää, T., & Pasanen, T. 2020. "Citizen Science in Information Systems Research: Evidence From a Systematic Literature Review," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, (doi:10.24251/HICSS.2020.569)

McGinnes, S. 2011. "Conceptual Modelling for Web Information Systems: What Semantics Can be Shared?," in *Advances in Conceptual Modeling. Recent Developments and New Directions*, O. De Troyer, C. Bauzer Medeiros, R. Billen, P. Hallot, A. Simitsis, & H. Van Mingroot (eds.), Berlin, Heidelberg, p. 4-13. (doi:10.1007/978-3-642-24574-9_3)

Murphy, R., & Parsons, J. 2020. "Capturing the Forest or the Trees: Designing for Granularity in Data Crowdsourcing," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Wailea, HI, USA, p. 395-404. (doi:10.24251/HICSS.2020.049)

Nov, O., Arazy, O., & Anderson, D. 2014. "Scientists@home: What Drives the Quantity and Quality of Online Citizen Science Participation," *PLoS one,* (9:4). (doi:10.1371/journal.pone.0090375)

Okoli, C., & Schabram, K. 2010. "A Guide to Conducting a Systematic Literature Review of Information Systems Research," *Sprouts: Working Papers on Information Systems,* (10:26).

Sanders, L., & Stappers, P. J. 2014. *Convivial Toolbox: Generative Research for the Front End of Design*, Amsterdam, The Netherlands: BIS Publishers.

Thuan, N. H., Antunes, P., & Johnstone, D. 2016. "Factors Influencing the Decision to Crowdsource: A Systematic Literature Review," *Information Systems Frontiers,* (18:1), pp. 47-68.

Wang, R. Y., & Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems,* (12:4), pp. 5-33.

Zhao, Y., & Zhu, Q. 2014. "Evaluation on Crowdsourcing Research: Current Status and Future Direction," *Information Systems Frontiers,* (16:3), pp. 417-434. (doi:10.1007/s10796-012-9350-4)

Zheng, H., Li, D., & Hou, W. 2011. "Task Design, Motivation, and Participation in Crowdsourcing Contests," *International Journal of Electronic Commerce,* (15:4), pp. 57-88.