Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content

Shawn Ogunseye, Jeffrey Parsons

Faculty of Business Administration, Memorial University of Newfoundland,
St. John's, Newfoundland & Labrador, Canada
{osogunseye, jeffreyp}@mun.ca

Abstract. Conventional wisdom holds that expert contributors provide higher quality user-generated content (UGC) than novices. Using the cognitive construct of *selective attention*, we argue that this may not be the case in some crowd-sourcing UGC applications. We argue that crowdsourcing systems that seek participation mainly from contributors who are experienced or have high levels of proficiency in the crowdsourcing task will gather less diverse and therefore less repurposable data. We discuss the importance of the information diversity dimension of information quality for the use and repurposing of UGC and provide a theoretical basis for our position, with the goal of stimulating empirical research.

Keywords: User-Generated Content, Information Diversity, Information Quality, Repurposability, Crowdsourcing.

1 Introduction

The development of interactive web technologies allows organizations to access information from individuals outside, and not formally associated with, the organization. This external information is commonly known as user-generated content (UGC) – content that is voluntarily contributed by individuals external to organizations. Access to UGC is revolutionizing industry and research. UGC sourced through crowdsourcing systems – systems that enable "outsourcing a task to a 'crowd', rather than to a designated 'agent' … in the form of an open call" [1, p355] – have successfully been used in diverse contexts for understanding customers, developing new products, improving service quality, and supporting scientific research [2–5]. In this paper, UGC and crowdsourcing refer specifically to UGC from purpose-built *integrative crowdsourcing systems*[1] that "pool complementary input from the crowd" [6, p98], rather than passive UGC collected through applications such as social media.

When creating crowdsourcing systems, one important design decision sponsors[2] must make is determining the composition of an appropriate crowd [28]. This decision influences the other design decisions about crowdsourcing projects (i.e. system

[1] Crowdsourcing systems that gather distributed information for decision making [6]

[2] Owners (key design decision makers) of crowdsourcing and crowd-facing systems [17]

Cite as Ogunseye, Shawn, and Parsons, Jeffrey. "Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content." *International Conference on Advanced Information Systems Engineering*. Springer, Cham, 2018.

design, task design, and motivation of contributors). Because the quality of UGC to be collected is a concern, sponsors either require potential contributors to possess relevant knowledge of the crowdsourcing task or allow a broader spectrum of volunteers to be part of their crowds. Choosing the former implies implementing recruitment strategies that favor knowledgeable contributors and prevent less knowledgeable contributors from participating, such as training volunteers before they are allowed to participate [8, 9] and recruiting experienced contributors – people who have previously participated (or are presently participating) in a similar project [31].

By restricting participation in integrative crowdsourcing projects to trained or experienced contributors, sponsors seek to tap into contributors' proficiency and familiarity with the task to ensure high information quality [30, 31]. This practice is supported in practice and in the crowdsourcing literature. For example, Wiggins et al.'s [p17] survey of 128 citizen science crowdsourcing projects – which often are integrative crowdsourcing systems that engage citizens in data collection – reports that "several projects depend on personal knowledge of contributing individuals in order to feel comfortable with data quality". Likewise, [8] promotes a contributor selection strategy for "eliminating poorly performing individuals from the crowd" and identifying experts from volunteers "who consistently outperform the crowd". However, in this position paper, *we make the case against adopting strategies that restrict participation to only knowledgeable contributors*.

2 Information Quality and Repurposable UGC

Knowledge about the phenomena on which data are being collected is assumed to positively influence the key dimensions of information quality – information accuracy and information completeness. Information accuracy is defined as "the correctness in the mapping of stored information to the appropriate state in the real world that the information represents" [10, p. 203], while information completeness is the "degree to which all possible states relevant to the user population are represented in the stored information" [10, p. 203]. However, the literature contains several studies in which experts or knowledgeable contributors in the crowd have not provided more accurate information than novices. For example, three studies in an ecological context found that knowledgeable contributors did not provide more accurate data than non-experts [11–13]. Likewise, in an experiment in which participants were required to identify and provide information about sightings of flora and fauna, novices performed as well as knowledgeable contributors with respect to the study's task [9].

Similarly, even though Kallimanis et al. [13] showed that less knowledgeable contributors report less information than knowledgeable contributors based on the fitness criterion employed in their study, they also reported that less knowledgeable contributors provided more data about certain aspects of the tasks than knowledgeable contributors and made significantly more unanticipated discoveries. These findings are mostly congruent with Lukyanenko et al.'s field and lab experiments [9, 16], which showed that the conceptualization and design of a crowdsourcing system plays a role in the completeness of data provided by contributors with varying degrees of knowledge. In sum, empirical research offers evidence that knowledgeable contributors do not always provide more complete or more accurate information (i.e. higher quality information) than those with little or no domain knowledge.

While accuracy and completeness are pertinent dimensions of information quality, UGC needs to encompass diverse views and perspectives to sufficiently address the need for contributed data to be repurposable [17]. This repurposability requirement can only be met if crowdsourced data is “managed with multiple different fitness for use requirements in mind” [18 p.11]. That is, the design choices made for integrative crowdsourcing systems should also support information diversity – the “number of different dimensions” present in data [7 p214] – to ensure repurposability and reusability of data. The relevant dimensions of information quality for crowdsourced UGC thus go beyond accuracy and dataset completeness and include information diversity.

Information diversity is *the amount of distinct information in contributions about an entity to the amount of information available in the contributions*[3]*. The degree of diversity between two contributions A and B, each consisting of a set of attributes, is* $\frac{(A \cup B - A \cap B)}{A \cup B}$*. The higher the index, the more diverse both contributions are*[4]. Information diversity promotes discoveries as it enables different users and uses of data, which may lead to unanticipated insights [17]. Information diversity helps provide a better understanding of data points, as some contributors may give details about the data point where others do not. In addition, information diversity affords flexibility to project sponsors, as data requirements may change with new insight or because projects are commissioned without clearly defined hypotheses in mind. A richer, more robust dataset can better handle such changes than a highly constrained one.

Understandably, information diversity has not received a lot of attention in the information quality literature, which has mainly focused on the quality of information collected within organizations with tight control over their information inputs, processing and outputs, and with predetermined users and uses of resulting data. Within these traditional organizational settings, described in [17] as *closed information environments*, information diversity is sometimes considered undesirable and data management processes seek to minimize or eliminate it. Moreover, in the few cases where data diversity has been considered in the context of the repurposability of UGC, research has focused on system (or data acquisition instrument) design [17–19]. Less attention has been paid to the effect of the cognitive diversity (i.e. differences in experience and task proficiency) arising from the choice of target crowds on the diversity of data generated.

3 Theoretical Foundation for Information Quality in UGC

Generally speaking, humans manage limited cognitive resources in the face of a barrage of sensory experience by paying selective attention to relevant features that aid in identifying instances of a class, while irrelevant features (those not useful for predicting class membership) can be safely ignored. Even though everyone selectively attends to information to some extent, our use of selective attention only covers top-

[3] This definition have been slightly modified after publication

[4] This definition is easily extended to the case where A and B are sets of contributions.

Cite as Ogunseye, Shawn, and Parsons, Jeffrey. "Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content." *International Conference on Advanced Information Systems Engineering*. Springer, Cham, 2018.

down attention, i.e. “internal guidance of attention based on prior knowledge, willful plans, and current goals” [14, p509].

Although selective attention leads to efficient learning, it is accompanied by the cost of learned inattention to features that are not “diagnostic” in the present context [21, 22]. Training leads to selective attention to pertinent or diagnostic attributes [22, 24]. When members of a crowd have been trained, their reporting will most closely align to the information learned from their training, resulting in less diversity than would be present in data reported by members of an untrained crowd. This is particularly pronounced when the training provides specific rules for performing the task, as contributors will tend to rely on (and pay attention to) this explicit information above any implicit inference they may form themselves – a phenomenon known as salience bias [15].

Consider a citizen science scenario (adapted from [22]) where contributors who have been trained on how to identify rose bushes were requested to report their occurrences in a field of rose, cranberry and raspberry bushes. In addition, assume contributors through their training are able to distinguish rose bushes from the other bushes present in the field by the absence of berries. Their training is sufficient to ensure the data they report is accurate and complete as other attributes like the presence of thorns would not be diagnostic in this context where rose and raspberry bushes both have thorns. However, if in the future a user needs to repurpose the collected data to confirm the presence of cranberry bushes in the same field or estimate their number, the presence or absence of berries is no longer diagnostic as cranberry and raspberry bushes have red berries, and the presence of thorns becomes diagnostic as cranberry bushes do not have thorns. The data becomes inadequate requiring resources to repeat the data acquisition stage. This tendency for training to influence the information reported by contributors making contributions align with the training received while reducing their diversity thus affects repurposability and the ability to make discoveries.

Similarly, experience increases the tendency towards selective attention. The absence of the tendency for selective attention is “a developmental default” [23, 24]. Infants do not selectively attend to attributes of instances. They reason about entities by observing all the features of individual instances [20] and are, therefore, naturally comparable to novice contributors in an integrative crowdsourcing context [24, 25]. The tendency for selective attention thus forms with development to aid classification as a mechanism for coping with the deluge of information around us. For this reason, the capacity to classify is a distinguishing factor between adults and infants [20]. As experience increases, the tendency for selective attention increases correspondingly.

Knowledge of the crowdsourcing task acquired by contributors through training or experience will help them report mainly about attributes of instances they have been taught (or learned experientially) to be relevant to the task [26]; thus, they are expected to be less likely to attend to attributes irrelevant to the task than novices [27]. Ogunseye and Parsons [29] argue that knowledge therefore affects the accuracy and completeness of contributed data as knowledgeable contributors have an increased tendency to only focus on diagnostic attributes, ignoring changes to other attributes when they occur. In addition, knowledgeable contributors show more resistance to further learning [27], impeding their ability to make discoveries. We add here that since contributors with similar knowledge are expected to show similar levels of selective attention and contribute more homogeneous data than cognitively diverse contributors, knowledge (task proficiency and experience) will also reduce a crowd's capacity for information diversity.

4 Conclusion

As organizations continue to leverage the collective wisdom of crowds, interest in crowdsourced UGC will continue to grow. At the center of new discovery and insight from UGC based on integrative crowdsourcing tasks rather than selective crowdsourcing tasks is the ability of collected UGC to accommodate the different perspectives of multiple users. This desire for repurposable UGC places a new information diversity requirement on crowdsourced information that is largely absent from traditional IS environments, where the uses of data are usually predetermined and stable. In addition to traditional dimensions of information quality, we argue for the inclusion of the information diversity dimension as a necessary dimension for crowdsourced UGC. We also explain from a cognitive perspective why training and experience will constrain information diversity and correspondingly, reduce the quality of crowdsourced UGC. Consequently, systems that seek repurposable UGC are better served if they are designed with inclusivity and openness as their core focus. Our agenda for future research includes studying how cognitive diversity impacts information diversity in different settings and how this impact affects the quality of decisions made from UGC.

References

1. Afuah, A., Tucci, C.L.: Crowdsourcing as a solution to distant search. Acad. Manage. Rev. 37, 355–375 (2012).
2. Castriotta, M., Di Guardo, M.C.: Open Innovation and Crowdsourcing: The Case of Mulino Bianco. In: Information Technology and Innovation Trends in Organizations. pp. 407–414. Springer (2011).
3. Hosseini, M., Phalp, K., Taylor, J., Ali, R.: The four pillars of crowdsourcing: A reference model. In: Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on. pp. 1–12. IEEE (2014).
4. Tarrell, A., Tahmasbi, N., Kocsis, D., Tripathi, A., Pedersen, J., Xiong, J., Oh, O., de Vreede, G.-J.: Crowdsourcing: A snapshot of published research. Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15-17, (2013).
5. Tripathi, A., Tahmasbi, N., Khazanchi, D., Najjar, L.: Crowdsourcing typology: a review of is research and organizations. Proc. Midwest Assoc. Inf. Syst. MWAIS. (2014).
6. Schenk, E., Guittard, C.: Towards a characterization of crowdsourcing practices. J. Innov. Econ. Manag. 93–107 (2011).
7. Hwang MI, Lin JW.: Information dimension, information overload and decision quality. Journal of information science. Jun;22(3):213-8 (1999).
8. Budescu, D.V., Chen, E.: Identifying expertise to extract the wisdom of crowds. Manag. Sci. 61, 267–280 (2014).

Cite as Ogunseye, Shawn, and Parsons, Jeffrey. "Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content." *International Conference on Advanced Information Systems Engineering*. Springer, Cham, 2018.

9. Lukyanenko, R., Parsons, J., Wiersma, Y.F.: The IQ of the crowd: understanding and improving information quality in structured user-generated content. Inf. Syst. Res. 22, 669–689 (2014).
10. Nelson, R.R., Todd, P.A., Wixom, B.H.: Antecedents of information and system quality: an empirical examination within the context of data warehousing. J. Manag. Inf. Syst. 21, 199–235 (2005).
11. Austen, G.E., Bindemann, M., Griffiths, R.A., Roberts, D.L.: Species identification by experts and non-experts: comparing images from field guides. Sci. Rep. 6, (2016).
12. Bloniarz, D.V., Ryan, H.D.P.: The use of volunteer initiatives in conducting urban forest resource inventories. J. Arboric. 22, 75–82 (1996).
13. Kallimanis, A.S., Panitsa, M., Dimopoulos, P.: Quality of non-expert citizen science data collected for habitat type conservation status assessment in Natura 2000 protected areas. Sci. Rep. 7, (2017).
14. Katsuki F, Constantinidis C.: Bottom-up and top-down attention: Different processes and overlapping neural systems. The Neuroscientist. Oct;20(5):509-21 (2014).
15. Lee HC, Ba S, Li X, Stallaert J.: Salience Bias in Crowdsourcing Contests. Information Systems Research. (2018).
16. Lukyanenko, R., Parsons, J., Wiersma, Y.F.: The impact of conceptual modeling on dataset completeness: A field experiment. 35th Int. Conf. Inf. Syst. ICIS 2014. (2014).
17. Parsons, J., Wand, Y.: A foundation for open information environments. (2014).
18. Woodall, P.: The Data Repurposing Challenge: New Pressures from Data Analytics. J. Data Inf. Qual. JDIQ. 8, 11 (2017).
19. Castellanos, A., Castillo, A., Lukyanenko, R., Tremblay, M.: Repurposing organizational electronic documentation: Lessons from Case Management in Foster Care. (2017).
20. Best, C.A., Yim, H., Sloutsky, V.M.: The cost of selective attention in category learning: Developmental differences between adults and infants. J. Exp. Child Psychol. 116, 105–119 (2013).
21. Colner, B., Rehder, B.: A new theory of classification and feature inference learning: An exemplar fragment model. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society. pp. 371–376 (2009).
22. Hoffman, A.B., Rehder, B.: The costs of supervised classification: The effect of learning task on conceptual flexibility. J. Exp. Psychol. Gen. 139, 319 (2010).
23. Gelman, S.A.: The development of induction within natural kind and artifact categories. Cognit. Psychol. 20, 65–95 (1988).
24. Kloos, H., Sloutsky, V.M.: What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. J. Exp. Psychol. Gen. 137, 52 (2008).
25. Keil, F.C.: Concepts, kinds, and conceptual development. Cambridge, MA: MIT Press (1989).
26. Harnad, S.: To cognize is to categorize: Cognition is categorization. Handb. Categ. Cogn. Sci. 20–45 (2005).
27. Plebanek, D.J., Sloutsky, V.M.: Costs of Selective Attention: When Children Notice What Adults Miss. Psychol. Sci. 956797617693005 (2017).
28. Malone, T.W., Laubacher, R., Dellarocas, C.: The collective intelligence genome. MIT Sloan Manag. Rev. 51, 21 (2010).
29. Ogunseye, S., Parsons, J.: Can Expertise Impair the Quality of Crowdsourced Data? Proc. 15th AIS SIGSAND Symp. Lubbock Tex. (2016)
30. Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K.: Mechanisms for Data Quality and Validation in Citizen Science. In: 2011 IEEE Seventh International Conference on e-Science Workshops. pp. 14–19 (2011).
31. Gura, T.: Citizen science: Amateur experts. Nature. 496, 259–261 (2013).

Cite as Ogunseye, Shawn, and Parsons, Jeffrey. "Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content." *International Conference on Advanced Information Systems Engineering*. Springer, Cham, 2018.