Contributor-centric Information Quality for Crowdsourcing

Position Statement

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Roman Lukyanenko
Memorial University of Newfoundland
roman.lukyanenko@mun.ca

Recently, there has been intense interest in *crowdsourcing* - wherein an organization calls upon the general public to carry out specific tasks in support of organizational objectives (see Doan et al., 2011). Applications of crowdsourcing are growing and include corporate product development, marketing research, public policy, and scientific research. Specially-built crowdsourcing platforms, such as Amazon Mechanical Turk or CrowdFlower, provide pools of 'crowdworkers' for hire.

Among other uses[1], crowdsourcing promises to dramatically expand organizational "sensor" networks, making it possible to collect large amounts of data from diverse audiences. As organizations increasingly employ crowdsourcing to collect information to support internal decision making, questions about the quality of information created by members of the crowd become critical (Sheppard et al., 2014, forthcoming; Antelio et al., 2012; Kremen et al., 2011; Arazy et al., 2011; Alabri and Hunter, 2010). Several approaches to information quality in crowdsourcing have been proposed (Lukyanenko et al., 2011; Wiggins et al., 2011). These include collaborative or peer review, leveraging redundancy in the crowds, and user training. Collaboration and peer review, for example, is the basis for iSpot (www.ispot.org.uk), a project that relies on social networking for collaborative identification of species of plants and animals (Silvertown, 2010). Crowd data can also be reviewed by experts (Sheppard et al., 2014, forthcoming; Hochachka et al., 2012). Whenever possible, organizations leverage redundancy in the crowds (e.g., by asking multiple observers to independently report on the same phenomena) (Franklin et al., 2011; Liu et al., 2012). Training is a common approach, especially when there are established standards to which contributions should adhere (Dickinson et al., 2010; Foster-Smith and Evans, 2003).

While these approaches appear promising, they have one general limitation: they define data quality from the perspective of the sponsoring organization and (potential) consumers of data. This focuses information quality improvement on issues related to qualifications and expertise of the contributors, aligning information capture with needs of those who may use the data. As, in most cases, data consumers want to organize information in a certain way (e.g., a biological taxonomy), this means that data contributors are asked to comply with some predefined schema. We argue this data consumer-centric approach may be limiting as it fails to fully account for the critical role of information producers in crowdsourcing settings.

While crowdsourcing projects are designed at the request of project sponsors – those who allocate resources (e.g., financial, management, and technical) to better serve the needs of (potential) data consumers - the data are produced by the members of the general public. There are often no constraints on who can participate and data are often produced by largely anonymous contributors with varying levels of domain expertise or motivation. Participants in these projects may not be aware or fully understand what project sponsors require or may not be willing to comply with sponsors' needs. For example, in citizen science, the views or biologists who intend to use data may be fundamentally different from views

[1] Other uses of crowdsourcing may include problem-solving, ideas generation, data analysis, creation of various artifacts (e.g., software, paintings, videos) (see, e.g., Doan et al., 2011).

of amateur participants with low levels of domain expertise (Parsons et al., 2011). This may result in low quality information, as non-expert contributors struggle to populate consumer-defined schemas with meaningful data (e.g., classify an observed bird as a biological species from a predefined list).

Considering the challenges of crowdsourcing, we propose a *consumer-centric* perspective on information quality. Such an approach entails few assumptions about how the information should be organized by data consumers; instead it focuses on whether the information faithfully represents views of information contributors. Consequently, we see decisions about how to collect and store information as critical to improving information quality in crowdsourcing. In particular, we advocate for an instance-and-attribute, rather than the traditional class-based, approach to organizing crowdsourced data (Parsons et al., 2011; Lukyanenko and Parsons, 2012). Our approach is based on the instance-based data model (Parsons and Wand, 2000), under which contributors are not forced to categorize instances of interest using predefined classes defined by project sponsors. This relaxes the constraint for non-experts to understand and conform to a chosen schema. This, we argue, should lead to data that are easier for the crowds to provide, faithful in representing contributors' views and are still be useful to project sponsors.

By reconceptualizing information quality from the contributor perspective, we hope to motivate efforts to design systems in ways that are sensitive to data contributors' points of view.

References

Alabri, A. and Hunter, J. 2010. "Enhancing the Quality and Trust of Citizen Science Data," *IEEE Sixth International Conference on e-Science*, pp. 81-88.

Antelio, M., Esteves, M.G.P., Schneider, D. and de Souza, J.M. 2012. "Qualitocracy: A data quality collaborative framework applied to citizen science," *2012 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 931-936.

Arazy, O., Nov, O., Patterson, R. and Yeo, L. 2011. "Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict," *Journal of Management Information Systems*, (27:4), pp. 71-98.

Dickinson, J.L., Zuckerberg, B. and Bonter, D.N. 2010. "Citizen science as an ecological research tool: challenges and benefits," *Annual Review of Ecology, Evolution, and Systematics*, (41:1), pp. 112-149.

Doan, A., Ramakrishnan, R. and Halevy, A.Y. 2011. "Crowdsourcing systems on the World-Wide Web," C*ommunications of the ACM*, (54:4), pp. 86-96.

Foster-Smith, J. and Evans, S.M. (2003). "The value of marine ecological data collected by volunteers," *Biological Conservation*, (113:2), pp. 199-213.

Franklin, M.J., Kossmann, D., Kraska, T., Ramesh, S. and Xin, R. 2011. "CrowdDB: answering queries with crowdsourcing," *ACM SIGMOD International Conference on Management of Data*, pp. 61-72.

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W. and Kelling, S. 2012. "Data-intensive science applied to broad-scale citizen science," *Trends in Ecology & Evolution*, (27:2), pp. 130-137.

Kremen, C., Ullman, K.S. and Thorp, R.W. 2011. ", Evaluating the Quality of Citizen-Scientist Data on Pollinator Communities Evaluación de la Calidad de Datos de Comunidades de Polinizadores Tomados por Ciudadanos-Científicos" *Conservation Biology*.

Liu, X., Lu, M., Ooi, B.C., Shen, Y., Wu, S. and Zhang, M. 2012. "CDAS: A crowdsourcing data analytics system," *VLDB Endowment*, (5:10), pp. 1040-1051.

Lukyanenko, R. and Parsons, J. 2011. "Rethinking data quality as an outcome of conceptual modeling choices," *16th International Conference on Information Quality*, pp. 1-16.

Lukyanenko, R. and Parsons, J. 2012. "Conceptual modeling principles for crowdsourcing," P*roceedings of the 1st international workshop on Multimodal crowd sensing*, pp. 3-6.

Lukyanenko, R., Parsons, J. and Wiersma, Y. 2011. "Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd," in Jain, H., Sinha, A. and Vitharana, P. (eds.), *Service-Oriented Perspectives in Design Science Research*, Springer Berlin / Heidelberg.

Parsons, J., Lukyanenko, R. and Wiersma, Y. 2011. "Easier citizen science is better," *Nature*, (471:7336), pp. 37-37.

Parsons, J. and Wand, Y. 2000. "Emancipating Instances from the Tyranny of Classes in Information Modeling," *ACM Transactions on Database Systems*, (25:2), pp. 228–268.
Sheppard, S., Wiggins, A. and Terveen, L. 2014, forthcoming. "Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data," *Computer Supported Cooperative Work and Social Computing*.
Silvertown, J. 2010. "Taxonomy: include social networking," *Nature*, (467:7317), pp. 788-788.
Wiggins, A., Newman, G., Stevenson, R.D. and Crowston, K. 2011. "Mechanisms for Data Quality and Validation in Citizen Science," *IEEE e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference*, pp. 14-19.