Information Quality Research Challenge: Adapting Information Quality Principles to User-Generated Content

ROMAN LUKYANENKO, Florida International University
JEFFREY PARSONS, Memorial University of Newfoundland

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]: Human Factors

General Terms: Design, Information Quality

Additional Key Words and Phrases: User-generated content

ACM Reference Format:
Roman Lukyanenko and Jeffrey Parsons. 2015. Information quality research challenge: Adapting information quality principles to user-generated content. ACM J. Data Inf. Qual. 6, 1, Article 3 (March 2015), 3 pages.
DOI: http://dx.doi.org/10.1145/2723166

Traditionally, information quality (IQ) research assumes organizational settings in which information production (e.g., internal, by external organizations/customers) is well-controlled and serves well-defined purposes. IQ research draws extensively on the manufacturing paradigm, treating information as a product and its quality as the extent to which information at hand fits consumer requirements [Ballou et al. 1998; Talburt et al. 2014; Wang and Strong 1996]. In a typical organizational environment, users who create data, professionals who curate it, and users who consume it are encouraged to closely coordinate activities [Lee and Strong 2003].

Although information production in corporate settings remains vitally important, organizations increasingly seek to harness data created beyond their boundaries. Of particular interest is User-Generated Content (UGC)—data produced by members of the general public rather than by employees or others closely associated with the organization. A major source of UGC is social networks (e.g., Facebook, Twitter); other sources include crowdsourcing (wherein users create content for specific purposes), product reviews, casual comments, tags, and annotated maps. In addition, a valuable complement of UGC is sensor data (e.g., user geolocations) transmitted by the devices (e.g., mobile, wearables) used to create data. Information produced by people with no formal links to an organization is a new, increasingly important, but poorly understood, addition to the IQ landscape.

Illustrating the potential of UGC, data production within organizations is being dwarfed by the amount created by ordinary people [Vellante 2010]. Companies are using UGC to understand customers and develop better products. Brynjolfsson and McAfee [2014] regard UGC to be one of the four pillars of the modern economy (along

Authors' addresses: R. Lukyanenko, College of Business, Florida International University, Miami, FL, United States, 33199; email: roman.lukyanenko@fiu.edu; J. Parsons, Faculty of Business Administration, Memorial University of Newfoundland, St. John's, NL, Canada, A1B 3X5; email: jeffreyp@mun.ca.

DOI: http://dx.doi.org/10.1145/2723166

with intellectual property and organizational and human capital). In the domain of citizen science (a tiny proportion of all UGC), Theobald et al. [2015] estimate the contribution of ordinary people to scientist-run crowdsourcing at $2.5 billion. These data promise to assist in tackling humanity's "evil quintet": climate change, overexploitation, invasive species, land use change, and pollution [Theobald et al. 2015].

Unlike typical corporate information production, it is difficult to control the content or form of data generated outside the organization. UGC is often accepted with little or no validation or controls. In addition, it is difficult to train and motivate casual content producers to provide high-quality data. Finally, the purposes for which UGC is produced may differ substantially from the way organizations intend to use it.

We believe the challenges of understanding and improving the quality of UGC warrant extending the prevailing principles and methods of IQ management to this domain. Whereas the organizational IQ paradigm focuses on the data consumer, IQ management of UGC should be more **contributor-focused**. Whereas training, instructions, and shared social norms ensure relatively transparent and controlled organizational information production, imposing such mechanisms on casual and open UGC is difficult. To properly understand the quality of data created by ordinary people, it is necessary to consider the motivation, abilities, and domain expertise of individual contributors. This may require IQ researchers to ground their work in relevant theories of human psychology.

Whereas organizationally-produced information is primarily intended for predefined uses, UGC is often **use-agnostic**. Casual content producers on social networks or crowdsourcing platforms may be unfamiliar with or unwilling/unequipped to satisfy the informational requirements of the organizations looking to use their data [Lukyanenko et al. 2014]. This is consistent with the general trend to mine datasets for unanticipated insights. To assist organizations in leveraging UGC, novel IQ metrics are needed that evaluate quality under the assumption of multiple, evolving, and unanticipated uses.

In cases where organizations sponsor UGC projects (e.g., crowdsourcing), a major question is how to design collection mechanisms and effectively engage contributors. Because it might be difficult to anticipate the structure (if any) of UGC, traditional mechanisms for collection (e.g., fixed forms) and storage (e.g., relational databases) may be inadequate (e.g., Lukyanenko et al. [2014]). Whereas traditional databases seek to minimize redundancy, in UGC settings redundancy (e.g., multiple users labeling the same item) may improve quality but raise novel problems of task modeling and cost-benefit tradeoffs (e.g., Ipeirotis et al. [2013]). Another challenge is motivation and compensation of contributors (e.g., Wang et al. [2012]). The heterogeneity of UGC also highlights the need for new approaches to information retrieval, record linkage and de-duplication, and data visualization to transform noisy and sparse data into forms amenable to organizational decision making. As more organizations begin to rely on information created by ordinary people, the need for novel conceptualizations and IQ management principles becomes ever more urgent.

REFERENCES

D. P. Ballou, R. Wang, H. Pazer, and G. K. Tayi. 1998. Modeling information manufacturing systems to determine information product quality. *Management Science* 44, 4, 462–484.

E. Brynjolfsson and A. McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* WW Norton & Company, New York, NY.

P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28, 2, 402–441.

Y. W. Lee and D. M. Strong. 2003. Knowing-why about data processes and data quality. *Journal of Management Information Systems* 20, 3, 13–39.

N. Levina and M. Arriaga. 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.

R. Lukyanenko, J. Parsons, and Y. Wiersma. 2014. The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research* 25, 4, 669–689.

A. Susarla, J. Oh, and Y. Tan. 2012. Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research* 23, 1, 23–41.

J. Talburt, T. L. Williams, T. C. Redman, and D. Becker. 2014. Information quality research challenge: Predicting and quantifying the impact of social issues on information quality programs. *Journal of Data and Information Quality* 5, 1–2, 1–3.

E. J. Theobald et al. 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation* 181, 236–244.

D. Vallente. 2010. Information explosion & cloud storage. Retrieved from http://wikibon.org/blog/cloud-storage.

J. Wang, A. Ghose, and P. Ipeirotis. 2012. Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? In *Proceedings of the International Conference on Information Systems*. 1–15.

R. Y. Wang and D. M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 4, 5–33.

Received November 2014; revised December 2014; accepted January 2015