

# MISQ Archivist

## Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content

*Roman Lukyanenko, Jeffrey Parsons,  
Yolanda F. Wiersma, and Mahed Maddah*

---

### Abstract

As crowdsourced user-generated content becomes an important source of data for organizations, a pressing question is how to ensure that data contributed by ordinary people outside of traditional organizational boundaries is of suitable quality to be useful for both known and unanticipated purposes. This research examines the impact of different information quality management strategies, and corresponding data collection design choices, on key dimensions of information quality in crowdsourced user-generated content. We conceptualize a contributor-centric information quality management approach focusing on instance-based data collection. We contrast it with the traditional consumer-centric fitness-for-use conceptualization of information quality that emphasizes class-based data collection. We present laboratory and field experiments conducted in a citizen science domain that demonstrate trade-offs between the quality dimensions of accuracy, completeness (including discoveries), and precision between the two information management approaches and their corresponding data collection designs. Specifically, we show that instance-based data collection results in higher accuracy, dataset completeness, and number of discoveries, but this comes at the expense of lower precision. We further validate the practical value of the instance-based approach by conducting an applicability check with potential data consumers (scientists, in our context of citizen science). In a follow-up study, we show, using human experts and supervised machine learning techniques, that substantial precision gains on instance-based data can be achieved with post-processing. We conclude by discussing the benefits and limitations of different information quality and data collection design choice for information quality in crowdsourced user-generated content.

**Keywords:** Crowdsourcing, user-generated content, citizen science, information systems design, information quality, information completeness, information accuracy, information precision, discovery, supervised machine learning