Crowdsourcing Approaches for Knowledge Organization Systems: Crowd collaboration or crowd work?

Maayan Zhitomirsky-Geffet
Bar-Ilan University, Israel
Maayan.Zhitomirsky-Geffet@biu.ac.il

Lala Hajibayova
Kent State University
lhajibay@kent.edu

Barbara H. Kwaśnik
Syracuse University, USA
bkwasnik@syr.edu

Juho Hamari
University of Tampere, Finland
jujohama@gmail.com

Julia Bullard
The University of Texas at Austin, USA
julia.a.bullard@gmail.com

Timothy Bowman
University of Turku, Finland
tim.bowman@gmail.com

ABSTRACT

Development of Internet technologies has empowered ordinary users to create, contribute, share and connect with other members of the community. As users learn to exploit the potential of networked communications, they participate in a process, which facilitates a shift from individual to collective contributions and introduces an opportunity for multi-vocal and multi-faceted representation of cultural heritage. Open access to crowdsourced collections requires reconsideration of the traditional authoritative approach of cultural heritage institutions. The arduous nature of the work rendered voluntarily in cultural heritage crowdsourcing initiatives calls for reconsideration of power relationships and giving power to devoted contributors supported by modern "intelligent" technology to regulate the process of representation and organization. Taking into consideration the fact that crowdsourced data are not without flaws, the question is how to better utilize the collective intelligence to create quality information. In this context, various issues such as power, control, trust, inter-contributor consensus, heterogeneity of opinions will be raised and discussed by the panelists. Each of the panelists comes from a different field of expertise (Computer science, Information science, Economics, Communication studies, cultural heritage) and various cultural backgrounds and geographical locations (United States, Europe and Israel). This diversity will be reflected in the presented perspectives on the crowdsourcing topic.

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

KEYWORDS: crowdsourcing, crowd collaboration, crowd work, ontologies, wisdom of crowds, collaborative knowledge organization

INTRODUCTION

Knowledge organization task involves extensive human expert participation/effort. In a recent review Simperl and Luczak-Rösch (2014) assert that today it is generally acknowledged that knowledge bases and ontologies should be developed and maintained in a community-driven manner, with tools providing collaboration platforms, enabling ontology stakeholders to exchange ideas and discuss modelling decisions. Hence, in the past years, numerous frameworks were proposed for collaborative ontology construction and integration by experts (Noy and Musen, 2003; Pereira, 2008; Shvaiko and Euzenat, 2013; Euzenat and Shvaiko, 2013; Simperl and Luczak-Rösch, 2014; Tudorache et al., 2008; Heflin, 2001; Holsapple and Joshi, 2002; Gómez-Gauchía et al., 2004; Karapiperis and Apostolou, 2006; Kalbasi, Janowicz, Reitsma, Boerboom, and Alesheikh, 2014).

The experts are able to build knowledge organization schemes and ontologies of high professional quality, but experts are hard to find and expensive to employ. Moreover, Muresan and Klavans (2013) argue that specialized medical terminologies, such as SNOMED CT cannot provide sufficient support when integrated into consumer-oriented applications because they do

not include lay vocabulary required for these applications.

Development of internet technologies has empowered ordinary users to create, contribute, share and connect with other members of the community (Jenkins et al., 2009). As users learn to exploit the potential of networked communications, they participate in a process that Lévy (2000) calls the collective intelligence, which facilitates a shift from individual to collective contributions and introduces an opportunity for multi-vocal and multi-faceted representation of cultural heritage. Cultural heritage institutions are traditionally considered authorities in the representation and organization of cultural heritage material, as well as in developing rules, standards and systems of representation and organization to facilitate accessibility and findability of the resources. Recently, however, the enormous growth of cultural heritage collections, as well as the pressure to make these collections accessible to the public, has become an underlying force motivating crowdsourcing projects.

Numerous digital photo archive and museum projects utilize the "wisdom of crowds" (Kotis and Papasalouros, 2011; Lin and Chen, 2012) to tag resources using their vocabulary. For example, Steve (Trant et al., 2007; Trant, 2009) and Powerhouse museums allow users freely assign tags to museum objects displayed on their websites. These projects allow users collectively tag objects from several museums and see both the metadata provided by the curators and the tags assigned by other users. Tag-based resource discovery system, utilized in these projects, facilitates retrieval of resources through tag cloud. Another crowdsourcing project is New York Public Library (NYPL) "What's on the menu" project that opens up its historic menu collection to the users to transcribe the menus, because many of the menus are handwritten or use fanciful typography and layouts that are not recognizable by mechanical translation methods ("What's on the menu?", 2015). Having searchable menu collection about dishes, prices and the organization of meals will help researchers to shed light on food history and culture.

The most crucial phase in the *crowd collaboration* process, which is not controlled or curated by experts, is for users with different opinions to reach a consensus. Three possible approaches to do this exist: 1) by intersection - where only the shared part of the different users' ontologies/classification schemes is included in the resulting ontology; 2) by union - where all the components constructed by different users are included in the final ontology; 3) by revision - where users might independently revise others' ontological components to reach a consensual ontology version.

There are many obstacles to creating harmonious collaboratively sourced multi-perspective knowledge representations. Understanding what the obstacles are at the fundamental level is a first step to creating processes and structures for acknowledging and leveraging such diversity. At the most basic level is the problem of identification and consensual definition of the components of such representations, especially when it concerns emerging forms and genres (Crowston, et al. 2011). Following that is the contextual nature of knowledge representation - for example, the representation of biological concepts in the fields of the science of Biology and in the practice of Forensics (Kwaśnik, 2011). Furthermore, representations that span different cultures and languages suffer from the same concerns as translation in general, where certain concepts may be missing or differently construed (Kwaśnik and Chun, 2004). Finally, even seemingly "universal" concepts, such as those expressing kinship, vary widely, even though the underlying phenomenon is understood more or less the same way (Kwaśnik and Rubin, 2003). Identifying a framework for accommodating (rather than masking or ignoring) such diversity will lead to structures that honor all contributions while at the same time maintaining the integrity of each.

An alternative approach for exploiting crowdsourcing without expert curation is by using technology for selection and regulation of knowledge organization schemes produced by *crowd work*. The underlying idea is recruiting a group of anonymous non-expert workers through some crowdsourcing site/s, such as Amazon Mechanical Turk or CrowdFlower for a given simple micro-task. Each worker is assigned a series of tasks, e.g. questions with a number of possible answers and is asked to select the most correct one in his/her opinion. To improve the quality of the replies, the workers on the site can be preliminarily tested on a small sample of qualification questions and only "the trusted workers" those who passed the test are selected for the real/main experiment.

As opposed to collaborative social crowdsourcing, in *technologically-regulated crowd work* the workers

independently fulfill a series of simple tasks, i.e. they do not directly interact during the work process, and thus, there is no need for reaching an inter-worker consensus in this approach. Rather, some automatic *aggregation measures/algorithms* are applied to calculate the collective decision out of the independent workers' answers. Such aggregated collective decisions have been shown to be as good as the domain expert's answers in various domains including economics, healthcare and cultural heritage, and for a much lower price (Howe, 2006; Quinn & Bederson, 2011; Cooper et al., 2010). Recently, Aroyo and Welty (2013) suggested that disagreement in crowd votes in crowdsourcing-based classification indicates vagueness or ambiguity in a sentence or in the relations being extracted. Numerous recent works have effectively utilized micro-task crowdsourcing technique for ontology construction, error detection and verification (Noy, Mortensen, Alexander and Musen; Mortensen, Musen, and Noy, 2013; Mortensen et al., 2015; Zhitomirsky-Geffet, Erez & Bar-Ilan, 2016).

Open access to crowdsourced collections requires reconsideration of the traditional authoritative approach of cultural heritage institutions. Taking into consideration the fact that crowdsourced data are not without flaws (e.g., incomplete information and spelling errors, just to name few), the question is how better to utilize the collective intelligence to provide access to quality information. Considering the fundamental concern of knowledge representation and organization systems to provide a user with the "best textual means to his end" (Wilson, 1968, p. 21), it is time to reevaluate the role of cultural institutions in crowdsourcing projects. In this regard, the power that is exercised through the panoptical design (Foucault, 1979) of, for example, the NYPL's "What's on the menu?" project allows the crowd only to transcribe and edit the given data while reserving exclusive authority to representatives of the library to monitor the whole process. According to Foucault's conceptualization of power, the pervasiveness of the panoptical system in our culture subjects all members to surveillance and relegates them to the position of the docile gazed-upon. Crowdsourcing, however, may represent a challenge to this system of social regulation. The arduous nature of the work rendered voluntarily in cultural heritage crowdsourcing initiatives calls for reconsideration of power relationships and a shift from authorized to multi-vocal distribution of the task, giving power to devoted contributors supported by modern "intelligent" technology to regulate the process of representation and organization.

STRUCTURE OF THE PANEL

The panel will begin with the moderator's brief introduction of the key issues with regard to crowdsourcing approaches to knowledge organization. Then, the following main issues will be discussed: 1) How to better utilize the collective intelligence to generate quality information, 2) What are the prominent approaches and methodologies for effective crowd collaboration, 3) What is the conceptual and practical difference and similarity between social crowd collaboration and crowd work? 4) Whether and how can non-experts collaboratively contribute to knowledge organization in various domains, such as cultural heritage, digital humanities, economics, communication and social studies. In this context, philosophical issues, such as power, control, trust, inter-contributor consensus, heterogeneity of opinions will be raised and discussed by the panelists in light of the existing theories. Each of the panelists comes from a different field of expertise and cultural background which will contribute to the diversity of presented perspectives. Each of the panelists will provide a five-minute overview of an aspect of crowdsourcing and knowledge organization. Further, the discussion will proceed among panelists, with contributions from the audience. Finally, a five-minute conclusion will be provided by the moderator, summarizing key ideas and questions for further research.

PANELISTS AND TOPICS

Maayan Zhitomirsky-Geffet

Dr. Zhitomirsky-Geffet received her PhD in Computer Science in 2006. Currently, she is an Assistant Professor in the Information Science department in Bar-Ilan University, Israel. Her primary field of research is ontology construction and development of methodologies and tools for collaborative knowledge organization and crowdsourcing technologies. In the past decade she published several articles on these topics in respected journals such as Journal of the Association for Information Science and Technology, Computational Linguistics, PLoS One.

In the proposed panel Dr. Zhitomirsky-Geffet will contribute the technological perspective of crowdsourcing approaches and their application for ontology construction and classification.

Lala Hajibayova

Dr. Lala Hajibayova is an Assistant Professor in the School of Library and Information Science at the Kent State University. Dr. Hajibayova received her Ph.D. in

Information Science in 2014 from Indiana University Bloomington. She holds a master's in Library Science from St. John's University, New York. Her research areas include knowledge representation and organization, metadata schemas, ontologies, information architecture, indigenous cultural heritage and computer-mediated communication. Dr. Hajibayova has published and presented at a number of national and international venues, including Journal of Information Science, Knowledge Organization, International Society for Knowledge Organization and Association for Information Science and Technology. She teaches in the area of knowledge representation and organization, and computer-mediated communication.

Dr. Hajibayova will talk about cultural heritage institutions crowdsourcing projects and issues of power, collaboration and trust.

Barbara H. Kwaśnik

Barbara H. Kwaśnik (MSLIS Queens College, CUNY, Ph.D. Rutgers University) is a professor at the i-School at Syracuse University, USA. Her research interests include organization of information, theory of classification, and information science. She is especially interested in how classifications intersect with everyday human endeavor - for example, how they are translated from one culture or application to another to help support increasingly diverse but interdependent contexts. Previous research (with Kevin Crowston) includes investigating whether genre information can help in searching, personal information management, and browsing.

Prof. Kwaśnik will discuss the challenges and processes of creating harmonious collaboratively sourced multi-perspective knowledge representations.

Timothy Bowman

Dr. Bowman's research interests include social theory as applied to online communications, scholarly communication, scientometrics/altmetrics, social media, social informatics, big data, and dynamic web application development. In essence, he is studying the intersection of information, communication, and technology at both the macro and micro levels. He is currently focused on applying social theory to study scholarly communication in online environments using large-scale datasets; the data that He is collecting, managing, and analyzing includes millions of records in combination with large bibliographic databases including the Web of Science, arXiv, and PubMed. Dr. Bowman's doctoral dissertation research examined the ways in which scholars frame tweets as personal or professional through the use of affordances. More specifically, he utilized social and ecological psychology theories from Erving Goffman (1959; 1974) and James J. Gibson (1977) to interpret the tweets of scholars that had been classified as personal or professional by workers (Turkers) on Amazon's Mechanical Turk (AMT) platform. He is also currently working on projects that include the study of scholars' presentation of self through profile images, whether altmetrics are an indication of scholarly impact, how scholars make use of social reference manager tools (e.g., Mendeley, Research Gate, etc.), and the need for applying social theory to understand scholarly communication via altmetrics.

Dr. Bowman will discuss societal impact of crowdsourcing.

Juho Hamari (Doctor of Economics) is currently a Principal Researcher at the Game Research Lab, School of Information Sciences, University of Tampere. Prior to joining University of Tampere he was a doctoral student and a researcher at Aalto University School of Business (2010-2016) as well as a researcher at Helsinki Institute for Information Technology HIIT (2008-2012). Dr. Hamari has also worked as Visiting Scholar at UC Berkeley School of Information (2015-2016). Dr. Hamari's research covers several forms of information technologies such as games, motivational information systems (e.g. gamification, game-based learning, persuasive technologies), new media (social networking services, online video streaming, eSports), peer-to-peer economies (sharing economy, collaborative consumption, crowdsourcing), and virtual economies. His research has been published in variety of respected journals such as Journal of the Association for Information Science and Technology, International Journal of Information Management, Computers in Human Behavior, Cyberpsychology, Behavior and Social Networking, Electronic Commerce Research and Applications, Simulation & Gaming as well as in books published by e.g. MIT Press.

Dr. Hamari will discuss issues related to shared economy and gamification.

Julia Bullard

Julia Bullard is a Doctoral Candidate at the University of Texas at Austin iSchool. She studies knowledge organization system design and has a particular interest in systems built by their users. Her dissertation work is an ethnography of a collaboratively designed organizing system for a large and growing fanfiction collection maintained by hundreds of volunteers from its community.

Julia brings to this panel, perspectives from knowledge organization, infrastructure studies, and values-in-design.

Acknowledgement

The panel is sponsored by SIG/Classification Research.

REFERENCES

Aroyo, L., & Welty, C. (2013). Measuring crowd truth for medical relation extraction. *AAAI2013 Fall Symposium on Semantics for Big Data*. Retrieved from https://www.aaai.org/ocs/index.php/FSS/FSS13/paper/view/7627/7543.

Chan, S. (2007), "Tagging and searching – Serendipity and museum collection databases", in *J. Trant and D. Bearman (eds.). Museums and the Web 2007: Proceedings, Toronto: Archives & Museum Informatics*, available at: http://www.archimuse.com/mw2007/papers/chan/chan.html (accessed 4 June 2011).

Cooper, S., Khatib, F., Treuille, A., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466, 756–60.

Crowston, Kevin, Kwaśnik, Barbara H., and Rubleske, Joe. (2011). Problems in the Use-Centered Development of a Taxonomy of Web Genres. Chapter 4. In: Mehler, Alexander, Sharoff, Serge, and Santini, Marina (Eds.) *Genres on the Web: Computational Models and Empirical Studies.* Springer.

Euzenat J. & Shvaiko P. (2013). *Ontology matching*, Heidelberg, Germany:Springer.

Foucault, M. (1979). Discipline and Punish: The birth of the prison. New York, NY: Vintage Books.

Gómez-Gauchía, H., Díaz-Agudo, B., & González-Calero, P. (2008). Two-layered approach to knowledge representation using conceptual maps and description logics. In: *Proceedings of the International Semantic Web conference (ISWC 2008). Lecture Notes in Computer Science*, 5318, 17-32.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, 5, 199-220.

Heflin, J. (2001). *Towards the Semantic web: Knowledge representation in a dynamic, distributed environment*. Unpublished doctoral dissertation, University of Maryland, College Park. Retrieved from http://www.cse.lehigh.edu/~heflin/pubs/heflin-thesis-orig.pdf

Holsapple, C. W. & Joshi, K. D. (2002). Ontology applications and design: A collaborative approach to ontology design. *Communications of the ACM*, 45(2), 42-47.

Howe, J. (2006). The rise of crowdsourcing. *Wired Mag*azine 14, 1–4.

Jenkins, H., Purushotma, R., Weigel, M., Clinton, K., & Robinson, A. (2009). Confronting the challenges of participatory culture: Media education for the 21st century. Cambridge, MA: The MIT Press.

Kalbasi, R., Janowicz, K., Reitsma, F., Boerboom, L., & Alesheikh, A., (2014). Collaborative ontology development for the geosciences. *Transactions in GIS* 18(6), 834–851.

Karapiperis, S., & Apostolou, D. (2006). Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*. 1(3), 199-216.

Kotis, K., & Vouros, G. A. (2006). Human-centered ontology engineering: The HCOME methodology. *Knowledge and Information Systems* 10(1), 109-131.

Kotis, K., & Papasalouros, A. (2011). Automated learning of social ontologies. In: Eds. W. Wong, W. Liu; M. Bennamoun *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, (pp. 227-246). IGI-Global.

Kwaśnik, B.H. (2011). Approaches to providing context in knowledge representation structures. In: *Proceedings. Classification and Ontology: Formal Approaches and Access to Knowledge*, UDC International Seminar 2011, The Hague, 19-20 Sept. 2011.

Kwaśnik, B.H. & Chun, You-Lee. (2004). Translation of classifications: Issues and solutions as exemplified in the *Korean Decimal Classification*. *Proceedings of the ISKO Conference*, London, England, July 2004.

Kwaśnik, Barbara H. and Rubin, Victoria Locktionova. (2003). Stretching conceptual structures in classifications across languages and cultures *Cataloging & Classification Quarterly*, 37, 1/2 (2003): 33-47.

Lévy, P. (2000). Collective intelligence: Man's emerging world in cyberspace. New York, NY: Perseus.

Lin, C. S., & n Chen, Y-F. (2012). Examining social tagging behavior and the construction of an online folksonomy from the perspectives of cultural capital and social capital. *Journal of Information Science* 38(6), 540-557.

Liu, J., & Gruen, D. M. (2008). Between ontology and folksonomy: A study of collaborative and implicit ontology evolution. In: *Proceedings of the 13th International Conference on Intelligent User Interfaces* (pp. 361-364). ACM, Gran Canaria, Canary Islands, Spain.

Mortensen, J. M., Minty E. P., Januszuk, M., Sweeney, T. E., Rector, A. L, Noy, N. F., & Musen, M. A. (2015). Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *Journal of American Medical Information Association.* 22(3), 640-648

Mortensen, J. M., Musen, M. A., & Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In: *Proceedings of the AMIA 2013 Annual Symposium* (pp. 1020-1029).

Muresan, S. & Klavans, J. L. (2013). Inducing terminologies from text: A case study for the consumer health domain. *Journal of the Association for Information Science and Technology*, 64, 727–744.

Noy, N. F. & Musen, M. A. (2003). The PROMPT suite: interactive tools for ontology merging and mapping, *International Journal of Human Computer Studies*, 59(6), 983–1024.

Noy, N. F., Mortensen, J. M., Alexander, P. R., & Musen. M. A. (2013). Mechanical Turk as an ontology engineer? Using microtasks as a component of an ontology-engineering workflow, In: *Proceedings of the 5th ACM Web Science 2013 Conference*, (pp. 262-271).

Pereira, C. S. (2008). Collaborative ontology specification. *Doctoral Symposium on Informatics Engineering*, Porto, Portugal. Retrieved from http://paginas.fe.up.pt/~prodei/DSIE08/papers/38.pdf

Quinn, A. J., & Bederson, B. B. (2011). Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the Annual Conference on Human factors in Computing Systems—CHI'11* (pp. 1403-1412). Vancouver, BC: ACM.

Sarasua, C., Simperl, E. & Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In: *Proceedings of the International Semantic Web Conference* (pp. 525-541). Boston, USA.

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* 25(1), 158-176.

Simperl, E., & Luczak-Rösch, M. (2014). Collaborative ontology engineering: a survey. *The Knowledge Engineering Review* 29, 101-131.

Smith B. et al. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, 25, 1251–1255.

Smith, B. (2008). Ontology (Science). In C. Eschenbach and M. Gruninger (eds.), Formal Ontology in Information Systems. In *Proceedings of FOIS 2008*, (pp. 21–35) Amsterdam/New York: ISO Press.

Trant, J. (2009), "Tagging, folksonomies and Art Museums: Early experiments and ongoing research". *Journal of Digital Information,10*(1) Retrieved from http://journals.tdl.org/jodi/article/viewArticle/270

Trant, J., Bearman, D. and Chun, S. (2007)." The eye of the beholder: steve.museum and social tagging of museum collections", in *International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings CD-ROM, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics*, Retrieved from http://www.archimuse.com/ichim07/papers/trant/trant.html.

Tudorache, T., Noy, N. F., Tu, S., & Musen, M. A. (2008). Supporting collaborative ontology development in Protégé. In: *Proceedings of the 7th International Semantic Web Conference (ISWC)* (pp. 17-32).

Voß, J. (2007), "Tagging, folksonomy & co - Renaissance of manual indexing?" Retrieved from http://arxiv.org/PS_cache/cs/pdf/0701/0701072v2.pdf.

Vrandecic, D., Vr, D., Pinto, S., Sure, Y., & Tempich, C. (2005). A diligent. The DILIGENT knowledge process. *Journal of Knowledge Management* 9(5), 85-96.

Weller, K. (2007), "Folksonomies and ontologies. Two new players in indexing and knowledge representation", in *Proceedings of Online Information 2007*. Retreived from http://wwwalt.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/35/1197280560weller009p.pdf.

Zhitomirsky-Geffet, M., Eden, E. S. and Bar-Ilan J. (2016). Towards Multi-viewpoint Ontology Construction by Collaboration of Non-experts and Crowdsourcing: The Case of the Effect of Diet on Health. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23686.