Crowdsourcing in the Digital Humanities: An Action Research on the Shengxuanhuai Manuscript Transcription

Yuxiang Zhao, Xuanhui Zhang, and Xiaokang Song

School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China
yxzhao@vip.163.com

Abstract. In recent years, there has been an emerging trend in the GLAMs (Galleries, Libraries, Archives and Museums) to leverage crowdsourcing to improve the collection, organization, and evaluation of valuable resources. Although a series of notable crowdsourcing projects in the digital humanities have been launched worldwide, there are few academic studies on investigating the implementation and evaluation of such cases. To fill up the research gap, this study aims at conducting a field exploration on the real case called the Shengxuanhuai Manuscript Transcription Initiative (Transcribe Sheng for short). In this poster, action research will be carried out to explore the various stages of Transcribe Sheng project. Our attempts may shed light on the design and evaluation principles of the crowdsourcing in the digital humanities.

Keywords: Crowdsourcing, Digital Humanities, Shengxuanhuai Manuscript Transcription, Action Research

1 Introduction

Crowdsourcing was defined by Howe as the "act of a company or institution taking a function once performed by employees and outsourcing it to an undefined network of people in the form of an open call" [1]. While research to date has mostly focused on crowdsourcing a task in the business context, in recent years, there has been an emerging progress in the GLAMs (Galleries, Libraries, Archives and Museums) to leverage crowdsourcing to improve the collection, organization, and evaluation of valuable resources. For instance, the national library of Australia recruited the public to correct the OCR of their digitized newspapers and achieved remarkable effects [2]. Crowdsourcing initiatives of GLAMs endeavor to offer citizens the opportunity to deeply involve in production, utilization, communication and curation of those feature collections and archives [3]. There is a growing trend within the digital humanities field to develop platforms and tools which outsource the traditionally time-intensive tasks to the mass volunteers. Such kind of projects can harness the wisdom of crowds to contribute to the utilizing and understanding of primary digital resources. In such case, there have been some attempts to crowdsource a more complex task traditionally

assumed to be handled by academics [4]. However, crowdsourcing in cultural heritage usually entails a greater level of time, effort, and intellectual input from the public [5]. Furthermore, sustained participation and contribution are of great challenges for those projects since the public may lose their interest after their initial trials.

The Shengxuanhuai Manuscript Transcription Initiative (Transcribe Sheng for short) is a cultural heritage-oriented crowdsourcing project which engages researchers and the general public. Transcribe Sheng can be labelled as a participatory archival project that seek public engagement and contributions to generate, describe, and identify the feature collections from Shanghai Library. Sheng Xuanhuai, Wade-Giles romanization Sheng Hsüan-huai (1844-1916), a Chinese government official, representative of Westernization Movement and entrepreneur in the last years of the Qing dynasty (1644–1911/12), responsible for much of China's early industrialization [6]. All his life, Sheng paid great attention to preserve the documents, such as various manuscripts, letters, books, or even menus. The collections have a great historical significance and the value of long-term preservation.

Transcribe Sheng as one of the first large-scale crowdsourcing projects in the cultural heritage domain in China, is of great significance to the development of Chinese digital humanities. Transcribe Sheng allows experts, amateurs, and general publics to help fully explore and widen access to this precious material. At the heart of the project is a collaborative transcription platform. A beta-test of Transcribe Sheng was released to the public in 2016, and the project was formally launched in the early 2017. The project aims at recruiting the public to transcribe the records of Sheng from 1850 to 1936, with 175000 volumes (c. 100 million words) in total. This study aims at conducting a field study on the implementation and management of crowdsourcing initiative in the burgeoning domain of digital humanities. Action research will be carried out to investigate the various stages of Transcribe Sheng project.

2　Related Work

With the digitization of a large amount of information resources such as books, journals, pictures, and newspapers, etc., and the rapid development of digital GLAMs, the quantity and quality of descriptions produced by the internal professional staffs are increasingly limited given the growing collections and downsizing of institutions [7]. In such case, crowdsourcing has been employed in the emerging field of digital humanities for a variety of tasks and activities, such as archival metadata generation, provide annotations and tags, and transcription, etc. In fact, some researchers have indicated that the paradigm of crowdsourcing has a perfect fit with the aim and scope of cultural heritage domains [8].

As a series of notable crowdsourcing projects in the digital humanities launched in the past few years worldwide, such as Transcribe Bentham, Ghostsigns, Old weather, and Your Paintings etc., a few academic work has been conducted in various contexts [9-11]. Some researchers focus on the specific crowdsourcing projects in cultural heritage domain. For example, Causer and Terras examined the implementation of Transcribe Bentham project and its contribution to humanistic studies and the bur-

geoning field of digital humanities [4]. Furthermore, some researchers paid closer attention to the classification and generalization of crowdsourcing in the digital humanities field from a macroscopic perspective. For example, Oomen and Aroyo identified six types of crowdsourcing based on activity characteristics in the cultural heritage domain [5]. Bonney concluded three models of participation in citizen science and social engagement, namely contributory projects, collaborative projects, and co-created projects, which may strengthen our understanding towards the adoption and adaption of cultural heritage-oriented crowdsourcing tasks in various stages [12]. Besides, Dunn and Hedges summarized four characteristics of crowdsourcing in digital humanities [13]. In addition, serious games or gamification play a critical role in attracting the crowds to contribute to cultural heritage related-crowdsourcing projects, which is a research hotspot in recent years [14-15].

However, there are still some challenges of crowdsourcing in digital humanities worthy of further exploration. Some researchers indicate that the technological issues, such as linked data, semantic web techniques and linguistic approaches are not well addressed and employed to facilitate the design of crowdsourcing platforms [5]. In terms of engagement challenges, Terras advocates that for some specific cultural heritage oriented crowdsourcing projects, sponsors and organizers need to explore a particular way to identify and recruit the right group of participants with great interest and fundamental knowledge to entry, rather than inviting the members from an undefined group [16]. Thus, how to set up some targeted incentive mechanisms to engage the participants and foster a sense of belonging towards the crowdsourcing projects in digital humanities are worth further examination. In addition, some researchers suggest that communication problems may result in the malfunction of crowdsourcing projects, such as the misunderstanding between the institutions of GLAMs and the general public, or the weak integration between discussion and task interfaces that may lead to poor user experience and further waste community effort [17].

3 Research Design

3.1 Research Approach

Given our objective of conducting a crowdsourcing initiative that focus on the implementation and evaluation of Transcribe Sheng, we will select action research as our mode of inquiry. As an interventionist method, action research highlights the real-world setting in which the working hypotheses about the phenomenon of interest can be examined by the researchers [18]. In addition to generating knowledge through field study in a real-world setting, action researchers place great emphasis on the change as an important outcome [19]. It is a cyclic and multiphase process consisting of five iterative phases, i.e., diagnosing, action planning, action taking, evaluating, and specifying learning [20-21]. In our case, action research should be a collaborative process emerging from the practical concerns of groups of researchers and participants working on transcription of Shengxuanhuai manuscripts. And it is particular suitable for a long-period research which requires researchers become deeply involved with the subject's problem and environment. The method combining theoreti-

cal knowledge of researchers with the practical actions of subjects has already been employed in some crowdsourcing projects [22-23]. However, there are few, if any, studies on crowdsourcing have adopted the action research to develop and test the related principles in digital humanities domain. In this study, some specific research methods, such as in-depth face to face interview, focus group, survey, and field experiment will be combined to investigate our topics.

3.2 Research Process

Transcribe Sheng was formally launched in early 2017. The objective of this project was to leverage the crowdsourcing paradigm to complete the transcription of 100 million words of manuscript history material and annotate. Figure 1 shows a screenshot of the prototype of crowdsourcing platform. The process is iterative in its nature and thereby fits into the action research approach. We follow Susman and Evered's cyclical action research design [19], and divide the research process into five stages as shown in Figure 2. Each stage consists of the five phases and the specific tasks, as illustrated in Table 1. Furthermore, according to Bonney's 3C (contributory, collaborative, co-created) rules of public participation level [12], we believe that with the progress of this project, participants' involvements are also strengthened. First, metadata design, iterative prototype development, and manuscripts transcription may benefit from the mass contribution of a large number of participants. Second, the project will gradually open the crosscheck function as well as cooperation standards to collaboratively establish the implementation principles on crowdsourcing in digital humanities. Third, we will use the linked data for mining historical events and present some visualization work in digital humanities. Some open contests may be launched to provide the participants with opportunities to carry out some co-creation projects based on the transcribed material.
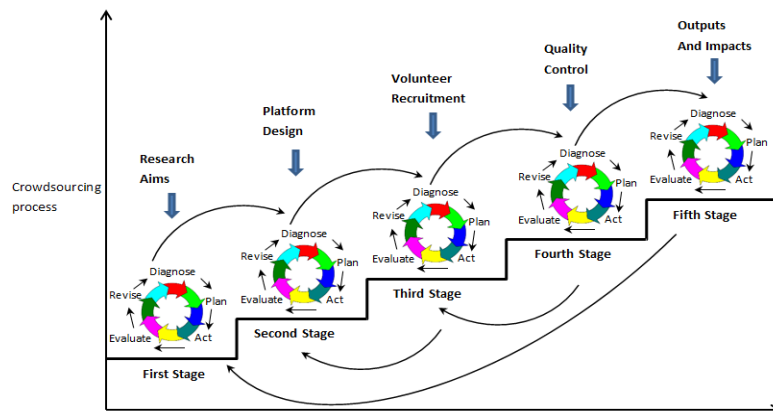

Fig. 1. The prototype of crowdsourcing platform

Fig. 2. Five stages of action research in the Transcribe Sheng crowdsourcing project

Table 1. Specific tasks in the various process of Transcribe Sheng

Process	Tasks	Participation
I	•To digitize 100 million words from Sheng Archives. •To annotate Sheng Archives with descriptive metadata.	Contributory
II	•To develop a prototype of online crowdsourcing platform for Transcribe Sheng project. •To iteratively design the usability and sociability of crowdsourcing platform.	Contributory
III	•To promote the project to the well selected communities of volunteer transcribers. •To transcribe the manuscripts in various ways.	Contributory
IV	•To crosscheck the content of transcription. •To improve data quality through expert's supervision.	Collaborative
V	•To use the linked data for mining historical events. •To use visualization tools to reproduce historical events. •To launch some open contests to facilitate the co-creation among the participants in digital humanities.	Co-created

References

1. Howe, J.: The rise of crowdsourcing. Wired Magazine 14(6), 1-4(2006).
2. Holley, R.: Crowdsourcing: How and Why Should Libraries Do It? D-Lib Magazine 16(3/4 Ma), (2010).
3. Owens, T.: Digital cultural heritage and the crowd. Curator the Museum Journal 56(1), 121-130(2013).
4. Causer, T., Terras, M.: Crowdsourcing Bentham: beyond the traditional boundaries of academic history. International Journal of Humanities and Arts Computing 8(1), 46-64(2014).
5. Oomen, J., Aroyo, L.: Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies, pp. 138-149. ACM, Brisbane (2011).

6. British Encyclopaedia, https://www.britannica.com/biography/Sheng-Xuanhuai, last accessed 2017/08/30.

7. Spindler, R. P.: An evaluation of crowdsourcing and participatory archives projects for archival description and transcription. Arizona State Univ. Libr 1-26 (2014).

8. Paraschakis, D.: Crowdsourcing cultural heritage metadata through social media gaming. Malmö University, (2013).

9. Moyle, M., Tonra, J., Wallace, V.: Manuscript transcription by crowdsourcing: Transcribe Bentham. Liber Quarterly 20, 3-4 (2011).

10. Carletti, L.: A grassroots initiative for digital preservation of ephemeral artefacts: the ghostsigns project. In: Digital Engagement Conference, PP. 15-17 (2011).

11. Ellis, A., Gluckman, D., Cooper, A., Greg, A.: Your paintings: A nation's oil paintings go online, tagged by the public. In: Museums and the Web, (2012).

12. Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., Wilderman, C. C.: Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. In: The Center for Advancement of Informal Science Education. Washington DC (2009).

13. Dunn, S., Hedges, M.: Crowd-sourcing scoping study: Engaging the crowd with humanities research. Centre for e-Research, King's College London. http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/12/Crowdsourcingconnected-communities.pdf (2012).

14. Flanagan, M., Punjasthitkul, S., Seidman, M., Kaufman, G., Carini, P.: Citizen archivists at play: game design for gathering metadata for cultural heritage institutions. Nature, 325(2013).

15. Ridge, M. Playing with difficult objects: game designs for crowdsourcing museum metadata. City University London, (2011).

16. Terras, M. M.: Crowdsourcing in the digital humanities. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) "A New Companion to Digital Humanities", PP. 420-439. Wiley-Blackwell, (2016).

17. Tinati, R., Van Kleek, M., Simperl, E., Luczak-Rösch, M., Simpson, R., & Shadbolt, N.: Designing for citizen data analysis: a cross-sectional case study of a multi-domain citizen science platform. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 4069-4078. ACM, Seoul (2015).

18. Lindgren, R., Henfridsson, O., Schultze, U.: Design principles for competence management systems: a synthesis of an action research study. MIS Quarterly 28(3), 435-472(2004).

19. Susman, G. I., Evered, R. D.: An assessment of the scientific merits of action research. Administrative Science Quarterly, 23(4), 582-603 (1978).

20. DeLuca, D., Kock, N.: Publishing information systems action research for a positivist audience. Communications of the Association for Information Systems 19(1), 10 (2007).

21. Davison, R., Martinsons, M. G., Kock, N.: Principles of canonical action research. Information Systems Journal 14(1), 65-86 (2004).

22. Haas, P., Blohm, I., Peters, C., Leimeister, J. M.: Modularization of Crowdfunding Services: Designing Disruptive Innovations in the Banking Industry. In: Thirty Sixth International Conference on Information Systems, Fort Worth (2015).

23. Leicht, N., Blohm, I., Leimeister, J. M.: How to Systematically Conduct Crowdsourced Software Testing? Insights from an Action Research Project. In: Thirty Seventh International Conference on Information Systems, Dublin (2016).