JADH 2017

Proceedings of the 7th Conference of Japanese Association for Digital Humanities

"Creating Data through Collaboration"

Japanese Association for Digital Humanities

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS

JADH2017

Proceedings of the 7th Conference of
Japanese Association for Digital Humanities

“Creating Data through Collaboration”

http://conf2017.jadh.org/
Doshisha University, September 11-12, 2017

Hosted by:
JADH2017 Organizing Committee
under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:
Department of Culture and Information Science, Doshisha University

Supported by:
Construction of a New Knowledge Base for Buddhist Studies:
Presentation of an Advanced Model for the Next Generation of Humanities Research
(15H05725, Masahiro Shimoda)
International Institute for Digital Humanities

Co-sponsored by:
PSJ SIG Computers and the Humanities
Japan Society for Digital Archive
Japanese Society for Information and Media Studies
Japan Art Documentation Society (JADS)
Japan Association for East Asian Text Processing (JAET)
Japan Association for English Corpus Studies
The Mathematical Linguistic Society of Japan
Japan Society of Information and Knowledge
Alliance of Digital Humanities Organizations

Creating Geotagged Humanities Data via Mobile Phone: Opportunities and Challenges

David Joseph Wrisley (New York University Abu Dhabi) Mario Hawat, Dalal Rahme (American University of Beirut)

Discussions of crowdsourcing in the digital humanities often concern textual data, in particular, transcription of manuscript data. Theorists point to the ability of "groups to out-perform individual experts, [and] outsiders [to] bring fresh insights to internal problems" (Brabham). In terms of data collection, mobile devices are described as freeing humans to collect significantly larger samples of data in space, in particular in the domains of citizen participation or infrastructure management. Crowdsourced data collection via mobile devices might seem like an ideal match, and yet the technique presents numerous challenges, foregrounding the necessity for digital humanities research to be vigilant about the changing modes of cultural knowledge production in "complex computational societies" (Berry/Fagerjord). Our paper discusses three ways that collecting data for digital humanities projects via mobile phones introduces new levels of data complexity: (1) the tension between ontological precision and on the fly human-intelligence tasking (2) the paradoxes of urban mapping research: redundancy, coverage and privacy (3) the impact of human behavior with socially embodied devices in sensitive environments.

Our paper discusses the experience of the "Linguistic Landscapes of Beirut" project (llbeirut.org). Our data consist of geotagged photos captured via mobile phones. The 2000+ images of urban multilingual writing were collected in two phases over two semesters in 2015- 16 year by a team of some thirty undergraduate researchers. They include metadata that are both automatically generated (time, latitude, longitude, image-size and phone model) and user-generated (language, script and linguistic features). A third post-processing phase of the project began in late 2016 focusing on more granular annotation and the transcription of the multilingual text found in the photos in YAML format.

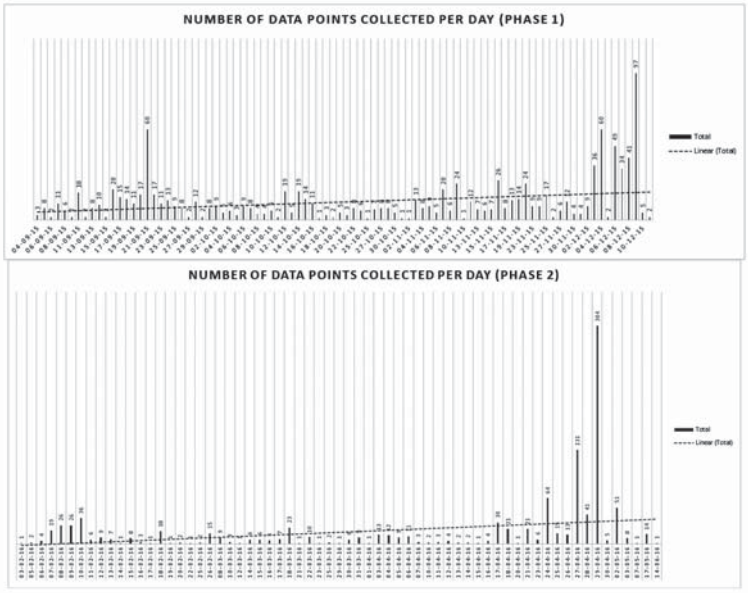
Our project corresponds to two aspects of Brabham's taxonomy of crowd participation: knowledge discovery and distributed human-intelligence tasking. Compared to other mobile application-based linguistic landscaping, our project is smaller in scale, yet richer in metadata (Lingscape). Whereas our data resemble volunteered geographical information (VGI), in that data collectors were free to capture images anywhere within an urban perimeter, it might be better described as a semi-directed, collaborative mapping since choice on the app data form was constrained to a bounded set of fields dictated by specific research questions.

First, an issue often discussed with VGI is the resultant data quality. Indeed, in the post-processing phase some inconsistencies in the classification of the samples were uncovered, but a more salient issue in the social creation of data was what we might call an ontological "shift," whereby the crowd avoided some categories and moved to nuance them in open comment fields. This was partially solved by iterative analysis of the data and collective reassessment of data fields. By keeping the number of "on the go" human-intelligence tasks to a minimum in the data entry form, we believed to be assuring data quality, but there were still a number of unclassifiable examples. We have attempted to deal with ambiguity in the post-processing phase via image annotation in order to qualify and identify these samples.

Second, whereas it has been argued that the geographical information from historical sources is inherently ambiguous and reflects user bias (Dossin et al.), the geolocation in our project was an automatic feature of the form builder. The data collection was left intentionally unstructured--participants were not obliged to use a specific sampling method. On the one hand, this meant that data accumulated along routes of the team's daily mobility. Again, iterative analysis of the data during the collection process showed the zones of greatest data density. Seeing this visualized in real time in-app and in web mapping environments encouraged some to venture

out into uncharted spaces in the city. On this point, we would argue that user specificity of the data reflected less a bias than contributing to the overall diversity of neighborhood coverage. This, however, required the anonymization of the dataset, since participants tended to collect data around their places of work and residence.

Third, discussions of VGI often mention user motivation. We did notice that although some "super users" (Causer and Wallace) emerged in the initial phases of the data collection, this abated, perhaps due to the social pressure not to over-perform in the pedagogical setting. Although pacing the data collection out over time was encouraged, the numbers of image samples captured tended to intensify at the end of each academic term as shown in Figures 1 and 2. We do not believe this to be a problem for the linguistic landscape data, but could have an impact on other projects that are more time sensitive.



Figures 1 and 2: A bar chart representing the number of data points collected per day over the two phases of data collection, with spikes at the end of each semester.

The mobile application method of the data collection "on the go" did allow us to scale up the data rather quickly, but it also revealed a cultural discomfort with photography in a divided city with highly visible security mechanisms. On the one hand, we realized that many of the photos were being taken out of car windows, a phenomenon for which the automatically-generated GPS_SPEED field provided some insight (Hawat). A more astonishing crowd pattern emerged however when we examined the aggregate of the data against the visible security mechanisms in municipal Beirut and found an uncanny avoidance of almost all of the secured zones of the city.

Figure 3: A map showing the aggregate of the "Linguistic Landscapes of Beirut" data (red) along with a rough depiction of the visible security mechanisms (yellow) following Harb et al.

This leads us to question what other blind spots exist in human behavior with mobile devices and encourages us to rethink the assumption of full freedom of movement through urban space when it comes to data collection.

Works Cited (all links accessed on 30 June 2017)

[1] Berry, D. and Fagerjord, A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*, Polity.

[2] Brabham, D. (2013). *Crowdsourcing*, MIT Press.

[3] Causer, T. and Wallace, V. (2012). Building A Volunteer Community: Results and Findings from Transcribe Bentham. *DHQ*, vol. 6.2 http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

[4] Dossin, C., N. Ningning Kong and B. Joyeux-Prunel. (2015). Applying VGI to collaborative research in the humanities: the case of ARTL@S, *Cartography and Geographic Information Science*. http://dx.doi.org/10.1080/15230406.2016.1216804

[5] Gilles, P. And Purschke, C. (2017). Lingscape. Mobile Application. https://lingscape.uni.lu/

[6] Hawat, M. (2016). "Mapping Mobility." Blog. http://www.mariohawat.com/blog/2016/5/16/moving- magic?p

[7] Harb, M., Gharbieh, A. and Fawaz, M., eds. (2009). *Beirut / Mapping Security*. IABR.

[8] March, J. (2016). "On Choosing a Mobile Platform in the Digital Humanities." Blog. https://digitalfellows.commons.gc.cuny.edu/2016/02/29/on-choosing-a-mobile-platform-in-the-digital-humanities/

Proceedings of JADH conference, vol. 2017
Published by the Faculty of Culture and Information Science, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto 610-0394 Japan
Online edition: ISSN 2432-3144　Print edition: ISSN 2432-3187
Editor: Akihiro Kawase