BIG DATA & SOCIETY

Big Data & Society
July–December 2014: 1–3
© The Author(s) 2014
DOI: 10.1177/2053951714544337
bds.sagepub.com
SAGE

Big Data and Small: Collaborations between ethnographers and data scientists

Heather Ford

Abstract

In the past three years, Heather Ford—an ethnographer and now a PhD student—has worked on ad hoc collaborative projects around Wikipedia sources with two data scientists from Minnesota, Dave Musicant and Shilad Sen. In this essay, she talks about how the three met, how they worked together, and what they gained from the experience. Three themes became apparent through their collaboration: that data scientists and ethnographers have much in common, that their skills are complementary, and that discovering the data together rather than compartmentalizing research activities was key to their success.

Keywords

Collaboration, interdisciplinary, ethnography, data science, Big Data, methods

In July 2011, at WikiSym in Mountain View, California, I met two computer scientists from Minnesota. I was working as an ethnographer for the non-profit technology company Ushahidi at the time, and I had worked with computer scientists before on tool building and design projects, but never on research. The three of us were introduced because we were all working on the subject of Wikipedia sources and citations.

We recently argued about who started the conversation. Dave Musicant, a computer scientist at Carleton College, said later that, although he loved doing interdisciplinary research, he was much too shy to have introduced himself. Shilad Sen is an Assistant Professor of Computer Science at Macalester College and had been working with Dave on a dataset of about 67 million source postings from about 3.5 million Wikipedia articles. In his usual generous manner, Shilad wrote later: "We had ground to a halt when you came to talk to us. We had done this Big Data analysis, but didn't have any idea what we should do with the data. You saved us!"

In retrospect, the collaboration that followed involved a great deal of mutual "saving." I was trying to build a portrait of how Wikipedians managed sources during breaking news events to inform Ushahidi's software development projects, but I did not have the bigger picture of Wikipedia sources to guide new directions in the research. Dave and Shilad were looking at whether one could predict which sources would stay on Wikipedia longer than others in order to build software tools to suggest citations to Wikipedians, but they had little detailed insight into why sources were added or removed in different contexts.

Over the next two years, the three of us met on Skype every few months to share our findings and then to conduct new analyses, test out new theories about the data, and finally produce a paper entitled "Getting to the source" (Ford et al., 2013) for WikiSym in 2013. I visited the two in Minnesota more recently to discuss the possible future trajectories for research, but our collaboration has remained informal and ad hoc. Despite this (or perhaps in large part because of this), my collaboration with Dave and Shilad has been one of the most enjoyable, educational experiences for me as an early career researcher. This is perhaps partly due to the unique

University of Oxford, Oxford, UK

Corresponding author:
Heather Ford, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK.
Email: heather.ford@oii.ox.ac.uk

combination of personalities that happen to combine particularly well, but I also think that interdisciplinary research like this can yield very exciting results for researchers coming from very different epistemological and methodological vantage points if they remain open and creative about the process. Three observations are particularly noteworthy here: that data scientists and ethnographers have much in common, that our skills are complementary, and that discovering the data together rather than compartmentalizing research activities was key to our success.

Ethnographers and data scientists have much in common

Although at first glance Big Data and ethnography can be seen in opposition (after all, ethnographers have their roots in studies of societies far removed from the heavily mediated ones of today), there are actually some significant commonalities. Both recognize that what people actually do (rather than only what they say) is invaluable, and both require an immersion in data in order to understand their research subject. As Jenna Burrell (2012) writes for *Ethnography Matters*:

> Ethnographers get at this the labor-intensive way, by hanging around and witnessing things first hand. Big data people do it a different way, by figuring out ways to capture actions in the moment, i.e. someone clicked on this link, set that preference, moved from this wireless access point to that one at a particular time.

Burrell argues that where there are differences is in the emphasis that ethnographers and data scientists place on what people do. Ethnographers, for example, do a lot of complementary work to connect apparent behavior to underlying meaning through in situ conversations or more formal interviews. Data scientists, on the other hand, tend to focus only on behavioral data traces.

If timed well, however, ethnographers and data scientists can come together at appropriate moments to collaborate on answers to common questions before moving on to wider (in the case of data science) or deeper (in the case of ethnography) research projects. In the case of the ‘‘Getting to the source’’ collaboration, the three of us shared a curiosity about sources and with Wikipedia practice more generally, and it was this shared curiosity that drove the project forward. I was interested in large-scale approaches to Wikipedia sources because I had been looking at Wikipedia's policy on sources and was finding in the examples and interviews that practice around sourcing was very different from what was being recommended in the policies. I was curious about whether source choices were, in fact, contradictory to policies that preferred academic sources. To understand whether my cases were indicative of larger trends, I needed to get a handle on the entire corpus of data traces. Shilad and Dave were interested in the ‘‘stickiness’’ of sources, trying to understand why some sources stuck around more than others. Sourcing practice, for them, was therefore really important for understanding how to analyze and evaluate the data traces represented in the database. All of us recognized the benefit of sharing skills and knowledge that we had gained in our different areas. I needed to understand ways of analyzing the entire corpus, and Dave and Shilad needed to understand everyday Wikipedia practice.

It turned out that, in addition to common questions and the need for shared expertise, we also shared commonalities in our approach. I was pleasantly surprised when I started working with Dave and Shilad that all of us preferred an approach that was inductive (testing out theories about the data as we progressed), systematic (being sure to follow up leads and challenge our assumptions), and collaborative (sharing responsibilities equally and understanding the decisions that we were all making and their impact on the project as a whole). I started this collaboration with an idea that quantitative research was largely deductive and that quantitative researchers would feel they had little to gain from working with those who tend to take a more qualitative approach. Through working with Dave and Shilad, however, I learned that we had much more in common than not, and that collaboration could yield worthwhile results for both data scientists and ethnographers.

Our skills and experience are complementary

In the Wikipedia research arena, a few Big Data researchers have used interviewing, participant observation, and coding in addition to their large-scale analyses to explore research questions. Brian Keegan's large-scale network analyses of traces through a system (Keegan et al., 2012) is an exemplar of Big Data research, for example, but Keegan also spent countless hours participating in the production of the class of Wikipedia articles that he was studying in order to understand the meaning of the traces that he was collecting. Keegan is, however, a rare example of an individual researcher who possesses the variety of skills necessary to answer some of the important questions of our age. More usual are the types of collaborations where researchers with a wide variety of skills and epistemologies work together to build rich perspectives on their research subjects and learn from one another in order to improve their skills and experience with methods with which they are unfamiliar.

In the case of the Wikipedia sources collaboration, Dave and Shilad had the necessary skills and resources to extract over 67 million source postings from about 3.5 million Wikipedia articles. Based on the interviews that I had done on ways in which Wikipedians chose and inscribed sources on the encyclopedia, I was able to contribute ideas about different ways of slicing the data in order to gain new insights. Dave and Shilad had access to sophisticated software and data processing tools for managing such a high volume of data, and I had the knowledge about Wikipedia practice that would inform some of the analyses that we chose to do on this data. After hearing from an expert interviewee that Wikipedians often discover their information using local sources but cite Western sources, for example, we were able to explore the diversity of sources in relation to their geographical provenance. By understanding this practice, we could also mention what was missing from the data that we had access to, namely, that citations did not necessarily represent what sources editors were using to find information, but rather what citations they believed others were more likely to respect. This small detail has significant implications for the conclusions that we draw about what sources and citations represent and the dynamics of collaboration on large peer production communities like Wikipedia. By discussing my findings with Dave and Shilad iteratively, we were able to come up with methods for operationalizing these hypotheses and developing different lenses for analyzing the data. Through this process, we recognized that our skills and experiences were highly complementary.

Discovering the data together is better than compartmentalizing activities

Where a large number of collaborative research activities fail is where tasks are divided up according to perceived skills and expertise of different types of research identities, rather than taking a more creative approach to research design. In this traditional view, ethnographers might be asked to do the interviews and manual coding where the Big Data analysts do the large-scale analyses with little collaboration and experience of these processes shared. The result is that there is no learning or sharing of skills: data scientists are seen merely as technicians who are able to manipulate the data and ethnographers as those who will "fill in" the context during write-up. If both researchers are to learn from the experience and stand on one another's shoulders to produce high-quality results, it is important that researchers share some unfamiliar tasks, or that they are at least taken through the processes that resulted in particular data being produced.

Although Dave and Shilad could have asked me to do the manual coding for our sources project alone, we decided to divide the tasks up so that we all contributed to the development of the coding scheme, and coded individual results and checked the accuracy of one another's coding. Although I led the development of a coding scheme, Dave and Shilad challenged me on the ways in which I was defining the scheme and both helped to manually code the random sample and to check my results. In this way, we all came out with a deeper understanding of the subject and of the ways in which our particular lens contributed to the shape of the research output. We also learned some important new skills. I learned how such large-scale analysis is done and about the choices that are made to achieve a particular result. Shilad, on the other hand, used the coding scheme that we developed together as an example in one of the method classes that he now teaches at Macalester College. We all extended ourselves through this project by sharing unfamiliar tasks and gaining a great deal more from this than we might have if we had kept to our traditional roles.

In summary, ethnographers have much to gain from analyzing large-scale data sources because they can provide a unique insight into how participants are interacting in complex media platforms in ways that complement observations in the field. Data scientists, in turn, can benefit from more qualitative insight into the implications of missing data, data incompleteness, and the social meanings attributed to data traces. Working together, ethnographers and data scientists can not only produce rigorous research but can also find ways of diversifying their skills as researchers. My experience with this project has given me new respect for quantitative research done well and has reiterated the fact that good research is good research whatever we call ourselves.

References

Burrell J (2012) The ethnographer's complete guide to big data: Answers. *Ethnography Matters*. Available at: www.ethnographymatters.net/blog/2012/06/11/the-ethnographers-complete-guide-to-big-data-part-ii-answers/ (accessed 9 July 2014).

Ford H, Sen S, Musicant DR, et al. (2013) Getting to the source: Where does Wikipedia get its information from? In: *Proceedings of the 9th international symposium on open collaboration*. New York, NY: ACM, pp. 9:1–9:10. doi:10.1145/2491055.2491064.

Keegan B, Gergle D and Contractor N (2012) Do editors or articles drive collaboration? Multilevel statistical network analysis of Wikipedia coauthorship. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. New York, NY: ACM, pp. 427–436. doi:10.1145/2145204.2145271.