Ethnography, Documents, and Big Data: Reflections on Teaching with David Hakken

Kalpana Shankar, University College Dublin
ei.dcu@raknahs.anaplak

Introduction

In the summer of 2011, I joined University College Dublin, Ireland, as a permanent faculty member in the then-named School of Information and Library Studies (now the School of Information and Communication Studies). As an academic working in a European university, I could avail myself of opportunities for European faculty and staff, including the Erasmus mobility program. The Erasmus program has provided teaching and training exchange opportunities for students, faculty, and staff between universities in Europe for over thirty years. David Hakken and Professor Vincenzo D'Andrea of the Department of Engineering and Computer Science at the University of Trento in northern Italy invited me to Trento to do an Erasmus exchange and co-teach with David, who was a former colleague of mine from Indiana University-Bloomington. David held a regular visiting fellowship with the University of Trento and spent several summers teaching and doing research there. During the summers of 2013 and 2014, I visited the University of Trento with funding from the Erasmus program to design and co-teach several workshops with David.

David was to teach a compressed research design class for graduate students in computer science and sociology both summers and wanted to integrate the workshops he and I planned into this class. We also wanted to open these workshops to academic staff and students in other departments to increase the visibility of these topics. After discussion about topics of mutual interest, we settled on two complementary topics: the ethnography of documents and the ethnography of big

data. David and I were both interested in ethnographic approaches to working with documents and data—not just as sources of information to support other methodological approaches, but also as objects of ethnographic study themselves.

Before I left for Trento in 2013, David and I finalized the topics of the workshops, sequenced them into the research design course he would be teaching, and drafted workshop descriptions and structures. We selected readings and supplementary materials and made them available to the students. We also had some idea of what the first workshop (the ethnography of documents) would look like and what hands-on activities would be required, so I found some open-access collections of documents. Once I arrived in Trento, David and I prepared each workshop together and divided the tasks for the actual teaching days.

In 2014, we reprised our workshops from the previous year but made some important modifications; some developments in the intervening year had caused us to rethink our approach to the big-data workshop. The well-documented failure of using Google Analytics to predict flu outbreaks (Butler 2013; Lazer et al 2014), the emergence of the *Big Data and Society* journal, new users of big data (citizen hackers and data journalists), and some more critical self-reflection from data scientists suggested to us that rather than an ethnography AND big data workshop (how can they be pursued together?), we would focus instead on ethnography OF big data (what IS big data?).

With two summers of teaching the workshops behind us, David and I considered how we might consolidate the work we had done for publication. We submitted a proposal to have a chapter on ethnography and documents included in the 4th edition of the *Handbook of Science and Technology Studies*, published by the MIT Press. The proposal was accepted after significant revision and with the support of an

additional co-author, Professor Carsten Oesterlund of Syracuse University. The final piece, entitled "Rethinking Documents," is in the 4th edition of the *Handbook of STS* (Shankar, Hakken, and Osterlund 2017); the chapter was finalized shortly before David's death and is dedicated to his memory.

The *Handbook* piece took a significantly different turn from what David and I had initially envisioned. This invitation to write for *Anthropology of Work Review* affords me the opportunity to revisit the teaching that David and I did over those two summers. Drawing upon the materials we created together, the conversations of the workshop participants, and subsequent reflections and discussions David and I had, I use this piece to describe our process and reflect on teaching and learning, primarily to computer scientists approaching sociotechnical research for the first time.

Studying Documentary Culture

The original anthropological ethnography, developed for study of peoples without writing, attended considerably to representational artifacts—from stories to dances to house poles—and their performance. This is because cultures depend on forms of representation, visual and oral but increasingly written. As ethnographic attention has turned to the more complex cultures of peoples for whom writing had become a privileged cultural practice, written documents' ethnographic centrality has increased. The study of "culture at a distance," primarily through written documents, that arose during World War II is arguably the first explicit form of documentary ethnography.

Many ethnographic sites are now heavily texted, often mediated by computing and studied through use of digital tools. Ethnographic emphasis on context means such large sections of research now depend on documents, whether computer code or

web pages, that require increasingly complex interpretation. Such digital representations mean documents are highly contingent and contextual not just in their content but also in format, organization, and relationships with other documents and artifacts.

Yet in our experience, many researchers are inclined to treat documents too close to face value. Anyone who has written an organizational minute, let alone more complex documents like professional codes of ethical conduct, can see that such assumptions are problematic. Yet they are taken as working assumptions for many current studies of big data; for example, that communication reception, via text messaging, means the existence of a relationship. (The message could have been received by accident.) Ephemeral documents like tweets are often used without considering their nature. These are just two examples of doubtful digital document research practices. Before we use them, or to use them better, we probably need to understand how documents came to be.

Workshop Development

In our ethnography of documents workshop, we wanted to accomplish two main tasks. The first was to introduce students to a wide range of scholarship on documents (and that there is such a thing!). We also wanted to awaken in our students an appreciation for the fact that documents pose numerous, insufficiently appreciated challenges for researchers. David and I surfaced enough pedagogical and research challenges that we felt we could easily build a workshop (or four) around the interrelated topics of studying documentary cultures and working with documents themselves.

We found it difficult to choose readings on the various ways of studying documents that have been extant since the late nineteenth century. At that time many American and European scholars worked together to create a number of international scientific associations, and many scholarly journals were founded, in large part to increase sharing of knowledge. The proliferation of scientific materials created a need for tools to locate colleagues' work, find publications, create and share collections of data, and coordinate collaborations. The notion of the document emerged as the organizing concept for a new "meta"-science of "documentalism," whose agenda concerned documents' material manifestations, temporal and spatial production, distribution, inventory, statistics, preservation, and use (Rayward 1994; Lund 2009). The study of the document, especially in organizational and other contexts, has continued unabated to this day (Riles 2006).

We also wanted to introduce research concerns that arise around collections and collectivities of documents: how should a researcher approach thinking beyond the single or even small group of documents, including archives? To tackle this challenge, scholars can learn methodologically and theoretically from studies of infrastructure and standards development, as well as studies of archives and archival practice. Anthropologists, science, technology, and society (STS) scholars, and archival-studies researchers have explored different ways that archives get created and "performed," including the connections of collecting practices for emergent social formations.

Present but mostly latent, these challenges have only grown. The essential point is that, instead of opportunistically taking documents for granted, we need to self-consciously deal with them in all their increasing complexity. They certainly are primary candidates for phenomena associated with contemporary life, such as the

production of knowledge and power and the dispersal of agency. Documents are evolving towards greater complexity, often via computer mediation. The ease of modification produces multiple iterations, whose changes offer insight into influence and power, but increased malleability also poses big problems regarding authenticity. Digital documents also pose challenges for maintenance and access over time. Ethnographic analyses depend heavily on placing practices in context, something increasingly composed of documents like algorithms or web pages. As digital representations, documents are highly contingent in format, organization, and relationships with other documents and artifacts, so understanding documentary contexts means mastering digital tools.

Ethical and legal challenges arise when studying documents. Dealing with privacy, anonymity, and content that is potentially incriminating (a recent internationally notorious parallel includes the Belfast Project, an oral history of members of the Irish Republican Army that implicated a sitting Irish politician in a decades-old murder) (George 2013). Other fields, such as archival studies, have dealt with similar challenges with respect to legal actions that are sought to return historical documents to places of origin (a concept called replevin) (Danielson 2013).

For the second workshop, the ethnography of data, especially big data, was not a great stretch. The behavioral meanings of findings from big-data analytics are arguably fraught and weighted with assumptions that are often "alien" to social research (Marres and Weltevrede 2013). David and I had had many conversations over the years around this issue of why quantitative analyses were presumed to be "epistemically neutral" and the role of ethnographic insight could take in either adding to or disrupting that train of analysis. We also talked about other big data topics and thought a workshop would allow us to raise some good questions. How

much and in what ways to technical affordances “matter”? What is the impact of analytic and interpretive tools on large data sets? All of these are issues that ethnography can potentially help address (Gross, Hakken, and True, 2012).

While big-data techniques can be used as a source of information regarding other topics—e.g., digital technology and social change—as part of methodological triangulation, the production of big data has made it a possible focus for contemporary cultural critique (Boyd and Crawford 2012). I had studied data cultures such as data repositories and archives and open government data but remained curious about what counted as big data. Both David and I felt that computer-science studies in particular take quantitative data for granted and would benefit from some exploration of big data as a problematic, while the sociology students would benefit from inquiry around issues of epistemic trust, power, agency, and discourse.

Structuring the workshops

For the Ethnography and Documents workshop, David and I divided the half day into three parts:

I. We and the attendees introduced ourselves and went over the plans for the day. David and I began with a presentation in which we discussed ethnography, the nature of the document, and how the two intertwine. We drew upon traditions from anthropology, information studies (particularly the documentalist tradition of the early mid-twentieth century and contemporary eras), and some current issues in research documents (such as open access, a topic we felt students need to be more familiar with). We discussed the nature of ethnography, the nature of

documents, the integration of the two, open access, history of the field, an overview of the readings, and issues related to studying the ethnography of documents.

II. We asked the students to separate into two teams. With guidance from David and me, they worked through a pre-sorted collection of documents to ask ethnographic questions (also prepared) about their structure, content, function, and the social realities they were reporting on.

III. Each team received a different set of documents. One team worked with a set of public health posters designed by the United States National Library of Medicine to raise awareness of the AIDS epidemic in the late 1980s. The other was a set of administrative documents from a Swiss pharmaceutical company. After a discussion about the documents with the questions below as a guide, we concluded the workshop with some final thoughts on ethical issues raised by working with and on documents. We asked the students to reflect on the following:

- What are the documents "evidence" of?
- What assumptions are embedded in the documents?
- What social relations are called upon to manufacture these documents?
- What rules of information selection, order, and viewpoint/voice are privileged in the documents?
- How are these rules applied in practice (in other words, how would you study the construction and role of the documents you have)?
- What else would you need to know in order to make ethnographic use of these documents (ex: what other documents, people, organizational structures, social networks, etc.).

- What methods would help you verify/question/critique whether your understanding of these documents is correct?
- What kind of research questions(s) could you design around these documents?
- What are the strengths and weaknesses of using these documents to answer your question(s)?

A week later, we held the second three-hour workshop, Ethnography and Big Data. This workshop drew more members of the larger University of Trento academic community who were interested in big data. We took a similar approach to the second workshop.

I. We introduced the participants, defined ethnography and big data, discussed mutual benefits to the two fields and challenges in rapprochement, discussed the readings, and presented the students with some documents to discuss as a group re: big data as a contemporary cultural phenomenon. We framed the discussion and the workshop as one of the possibilities of rapprochement or entrenched hostility (with each of us taking different views).

II. Attendees divided themselves into four teams. We asked them to develop an ethnographic agenda around big data. We provided them with some trade press and media articles on big data, including links to stories on Google Flu, economic aspects of the big-data economy, educational programs in data science and MOOCs, data journalism, smart cities, links to journals, and some pieces from the group blog Ethnography Matters. On their own, they were asked to find "evidence" of the themes of the workshop and discussion and formulate two "next steps for ethnographers" in studying big data. We provided some questions for the teams to begin with:

- Who are the big data "natives," and what do they think (and say) big data is?
- How could one study big-data practices and how do those practices relate to what the natives think?
- What do we think we know, if only provisionally, about what big data is, and what else do we need to know?

Discussion

The ethnography of documents workshop proved to be "easier"—both for us to articulate and for the attendees to engage. The experiential research methods of ethnography continue to attract people working with digital technologies, not only to inspire design but also as a basis for a deeper and longer engagement with the wide range of digitally mediated phenomena. Working with written records and other documents together suggested to the students that the ethnography of documents could be a fruitful venue of approach. The students suggested participant observation of the creation of particular genres of documents (policy, advertising, public-service announcements, to name a few) and triangulating observation with examining paper and digital documents could combine the insights acquired by "hanging out" (to use David's phrase) with recognition of how documents can be active agents. In particular, we discussed the relationship of digital documents (and computing more broadly) with massive social change, built deeply, for example, into much social policy. We were able to make more explicit the widespread belief that digitization is socially transformative.

Our experiences suggested that the following queries should become standard in research if they are not already: What documents are relevant to your research? If your documents are formal reports, what prior documents are called upon to manufacture those reports? What are the limits and boundaries set by the report? How

is “bad news” constructed? What rules of information selection, order, and viewpoint/voice are privileged in the documents? How are these performed/applied in practice? What assumptions are embedded in the documents? How do your documents mark and define social networks? What are the documents “evidence” of? What methods would help you verify/question/critique the documents? How do documents participate in a web of interactions with other artifacts? We finished our workshops aware that we could do much more with more time, but that there was much to say—which is what we did with the *Handbook* chapter.

The ethnography and/of big-data workshops (both of them) were more difficult for us to teach and for the students to engage. In some part, this was because David and I did not have a ready source of a hundred years of scholarship to build on as we did with document studies. Second, the topic to some extent annoyed us both. We were both academically affected by various dismissals of “our” work and studies of social phenomena as “unscientific” or “unnecessary” because big-data analyses of social-media data could supercede close qualitative work. We had both tried in various ways to work with data scientists and were rejected or, in my case, my work was co-opted without credit. Lastly, this nascent area of work was only just beginning to raise ethical issues around informed (or other) consent, access to knowledge, equity of access, and other areas of inquiry, and we were as yet unsure in ourselves how to address them. Doing ethnography of big data seemed interesting, but was it even possible? What would we do as ethnographers—sit in a cubicle farm watching people executing machine-learning algorithms?

As we were working through the workshops and the chapter we wrote (over glasses of prosecco and walks in the hills above Trento), we had questions for each other that we were never able to resolve. One important research question is what the

goal of document and data ethnography could and should be. Is the goal to get to ethnographic "reality" thru these "artifacts," or to specify directly their respective ethnographic realities? What is the relationship between data, documents, and other material culture items? When, for example, is document knowledge itself, when merely a representation thereof, and when does a document function as camouflage for "real" patterns of culture? In other words, how are we to comprehend the dialectic between the multiple formalities of data and document knowledge's substantial informality and embodiment?

And to go back to our starting point, how could we engage students and early-stage researchers in asking these questions? David and I were both interested in scholarly disciplines and the political ecology of universities. Our short workshop format and the different disciplines of our attendees forced us to raise questions about the institutional homes of document and data studies and their implications. To give one example, the iSchool movement, involving scholars from across the humanistic and scientific disciplines, was both forcing and fostering an encounter between disciplinary approaches to data and documents. David argued that even developments in fields more distant, such as literature departments' interest in "reader response theory," made document ethnography more necessary as STS finds its place in the academy. In my field of information studies, David pointed to the failure of the "knowledge management" program of the 1990s, for example, which can in part be traced to proponents' tendency to confuse an organization's knowledge with the representations of it, a set of documents being merely repositioned as a so-called "knowledge base" (Wilson 2002). He argued that such approaches alienate documents from their profound sociality.

Conclusion

In this piece, I have used the opportunity to finish an unfinished conversation. There was much more David and I could have said to each other, to our colleagues, and to our students. Toward the end of his life, David said to me that he felt that he was at the stage in his career where teaching and mentoring a new generation of scholars was the most important work he could do. A piece on teaching and learning about topics that he was passionate about seems like a fitting way to honor his legacy.

Notes

I have made the notes for our workshops available at my website: http://kalpanashankar.com/index.html on the Teaching and PhD Supervision tab. Please feel free to use them with attribution.

References

Boyd, Danah, and Kate Crawford.
2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15:662–679.

Butler, Declan.
2013. "When Google Got Flu Wrong." *Nature* 494:155–156.

Danielson, Elena S.
2013. "Archives and the Ethics of Replevin." *Journal of Information Ethics* 22:110–140.

George, Christine.
2013. "Archives Beyond the Pale: Negotiating Legal and Ethical Entanglements after the Belfast Project." *The American Archivist* 76:47–67.

Gross, Shad, David Hakken, and Nic True.
2012. "Studying Social Relations in MMOG Play: An Illustration of Using Ethnography to Frame 'Big Data'." In *IEEE Computer Games (CGAMES), 2012 17th International Conference on*. Pp. 167–174.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani.
2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343:1203–1205.

Lund, Niels Windfeld.
2009. "Document Theory." *Annual Review of Information Science and Technology* 43:1–55.

Marres, Noortje, and Esther Weltevrede.
2013. "Scraping the Social? Issues in Live Social Research." *Journal of Cultural Economy* 6:313–335.

Rayward, W. Boyd.
1994. "Visions of Xanadu: Paul Otlet (1868–1944) and Hypertext." *Journal of the American Society for Information Science* 45:235–50.

Riles, Annelise.
2006. *Documents: Artifacts of Modern Knowledge.* Ann Arbor: University of Michigan Press.

Shankar, Kalpana, David Hakken, and Carsten Osterlund.
2017. "Rethinking Documents." In *Handbook of Science and Technology Studies, Fourth Edition*. Ulrike Felt, Rayvon Fourché, Clark A. Miller, and Laurel Smith-Doerr, eds. Pp. 59–86. Cambridge, MA: The MIT Press.

Wilson, Tom D.
2002. "The Nonsense of Knowledge Management." *Information Research* 8,

no. 1 (2002). Accessed April 27, 2017. http://InformationR.net/ir/8-1/paper144.html.