Data-Sprinting: a Public Approach to Digital Research

Tommaso Venturini, Anders Munk, Axel Meunier

How to cite this paper:
Venturini, T., Munk, A., & Meunier, A. (2016). Data-Sprint: a Public Approach to Digital Research. (C. Lury, P. Clough, M. Michael, R. Fensham, S. Lammes, A. Last, & E. Uprichard, Eds.) *Interdisciplinary Research Methods* (forthcoming).

Abstract

This chapter is about the politics of interdisciplinarity. Not in the sense of the research politics fostering collaboration across disciplines, but in the stronger sense of transcending disciplinary boundaries to make significant political contributions. In short: it is about making research public. To address this question, this chapter introduces (through a concrete example in climate debate research) an original research format, that we call data-sprinting.

« It is controversies of this kind, the hardest controversies to disentangle, that the public is called in to judge. Where the facts are most obscure, where precedents are lacking, where novelty and confusion pervade everything, the public in all its unfitness is compelled to make its most important decisions » (Lippmann, 1925, p. 121).

What's in a data-sprint

Data-sprints are intensive research and coding workshops where participants coming from different academic and non-academic background convene physically to work together on a set of data and research questions.

Data-sprints have their roots in a series of organizational innovation introduced in the field of open-source development at the turn of the century (as a reaction to the previsous 'waterfall approach' inherited from the engineering management – Raymond, 2001). Faced with such radical uncertainty about how their project will develop and who will join them, open-source developers invented a form of coding event called "barcamps" or "hackathons" (or hacking marathons). Such format consists of short events in which a group of developers and designers meet to work intensively and expeditiously on some digital object.

The features of hackathons and barcamps fit extremely well the practice of interdisciplinary research. We appreciated in particular:

1. The heterogeneity of the actors involved. Hackathons and barcamps are generally organized as to be open to many different types of actors. In part, this comes from the need to achieve deliverable results at the end of the event, which demands to collect all the necessary competences to go through all the phases of the project. In developing marathons, this translates in having experts from the entire programming stack: from setting up the server infrastructure, to designing the wireframes, from scraping the data to implementing the front-end. The push for heterogeneity also derives from the necessity to exchange with the potential end users of the projects, who should be at hand during the developing dash.

2. The effort to convene participants physically. The unity of time and place that characterize hackathons and barcamps seemed an appropriate counterbalance to the dispersion of research efforts often observed in international and interdisciplinary projects. One of the problems of working across disciplines is that the experts of one field have a blurred appreciation of what experts of other fields may need as an input for their work. Such misunderstandings are normal in interdisciplinary projects and can become disastrous if discovered too late – a risk particularly salient for international projects. Yes, technologies for distant cooperative work can ease up some of these difficulties, but nothing facilitates mutual supervision and speed up collaboration more than direct presence. One more time, 'digital' turned out to be opposite to 'virtual'. Exploiting digital inscriptions demands to coordinate the efforts of many different disciplines and this in turn demands to convene in the same space and time.
3. The "quick and dirty" (or "design to cost") approach. Though thriving on the same increase in the availability of digital inscriptions, hackathons and barcamps are somewhat opposite to 'big data' approaches. The short and intensive nature of these events shields them from the dream of exhaustivity often associated with 'big data'. Participants know that they will only able to treat a limited amount of digital traces and that they will achieve imperfect results, but they accept such constraints more as a challenge than as a weakness. Making the most out of light infrastructures, simple logistics and agile organization methods, participants are well aware that their work should hack code and information gathered in earlier projects and that their outcomes should become the bases for further ventures. Not only hackathons and barcamps foster iterations within their meetings, but they are explicitly conceived as intermediary steps of a larger developing cycle.

With the format of data-sprint, we tried to adapt hackathons and barcamps to the practice of academic research by adding the larger efforts of 'contextualization' before, during and after the event:

1. Data sprints are always preceded by a long and intense work of preparation. When participants meet up, most of the research infrastructure should have already been collected and prepared for treatment. Time-consuming operations such as data cleansing or infrastructures setting-up should be accomplished beforehand, so that the days of the sprint can be dedicated entirely to the operations that require a more direct collaboration. Also participation to data-sprints is not open: sprinting lineup and team formation is taken care in advance, to make sure that the working groups contains all the competences needed to achieve significant results.
2. Data-sprints are also generally longer and more structured than their antecedents. While hackathons and barcamps are usually organized on two or three days, sprints work better when they extend over a full working week.
3. Finally, data-sprints require a greater follow-up than hackathons and barcamps. The 'quick end dirty' approach that characterizes the five days of a sprint should be complemented by an extensive work of refinement and documentation, in order to endow the results with the precision and robustness demanded by scientific research.

For the sake of clarity, it is possible to single out 6 different phases of a data-sprint that, though mingled in the practice of data-sprints (because of the flexible and iterative nature of such approach), correspond to distinctive organization concerns:

1) Posing research questions. Research questions are posed on the first day of the sprint by the invited issue experts. Besides suggesting a number of research questions, issue experts are also invited to help the

other participants (most of whom have little previous knowledge of the issue at stake) to get grips with the topic of the meeting. This can be done through Q&A sessions or panel discussions, but it also (and often more fruitfully) through informal consultations as part of the running feedback on data visualizations.

2) Operationalizing research questions into feasible digital methods projects. In a sense, this process begins already before the sprint where the organizers try to anticipate what type of projects the sprint might led to. We found that an excellent way of doing this initial vetting is to ask issue experts to suggest interesting datasets. This provides a chance to get back to the experts explaining why the proposed dataset may be unsuitable for certain research questions and thus getting them attuned to what a digital methods project can and cannot achieve.

3) Procuring and preparing datasets. As mentioned above, while it is desirable to have datasets available in advance, this is sometime at odds with the agility of the sprint and it is not uncommon that complementary data have to be search and collected in the first days of the sprint.

4) Writing and adapting code. Sprints are issue-specific (meant to address particular needs of the controversy actors) and their aim is less to develop generic tools than to adapt existing code to the research questions rose by the issue experts. This does not mean, however, that effort shouldn't be invested in making datasets, scripts and visualizations re-usable beyond their original projects. Sprints should remain faithful to their communitarian roots and ensure that all the data, code and contents produced are liberated through of open-source, copy-left and open-publishing licenses.

5) Designing data visualizations and interface. One of the driving forces of sprints is that they deliver tangible outcomes. These outcomes may have different forms, but they always share the characteristic of being directly usable by actors of the controversy. In many cases, this translates with the issue experts leaving the sprints with tangible results that they immediately mobilize in their debates.

6) Eliciting engagement and co-production of knowledge. Data-sprints abide by the 'co-production of knowledge model' of social sciences advocated for by Callon, Lascoumes, & Barthe (1999). Such approach assumes that scientific activities should be pursued in a constant and genuine dialogue with their publics. If data-sprints are organized according to the five phases described above, it is distinctively for this final phase to be achieved. If sprints fail in creating a common space for social scientists and social actors, they will have failed in all other respects as well.

EMAPS and the example of climate adaptation

To illustrate the research situation in which data-sprinting can be useful, we draw here on a concrete experience of a 3-year EU-funded collaborative project called EMAPS (Electronic Maps to Assist Public Science). EMAPS was a project in controversy mapping (Venturini, 2010 and 2012) with the specific objective to analyze the public debate about climate change *adaptation*. Discussions about how to cope with the impacts of climate change have become particularly salient in the last few years after the recurrent failures to reduce greenhouse gases (GHGs) emissions (Aykut & Dahan, 2015).

Adaptation constitutes one of the most intricate controversies of collective existence: actors enter and exit the discussion as recklessly as the rise and fall of issues; coalitions form and dissolve hectically; and conflicts crosscut each other making it difficult to identify opposing factions. In such overflowing complexity, existing institutions are so completely overrun by the shifting of alliances and oppositions that functionalist and

critical approaches lose much of their interest. In the debate on mitigation, investigating which international organizations are most suitable to regulate GHG emissions or which companies are most liable for them (Heed 2014) makes perfect sense. Not in the debate on adaptation. When it comes to imagining how to live through the radical changeover of global warming, distributing blame and praise is less important than working with actors to make new collective arrangements possible.

Yes, but what actors? Willing as we were, at the outset of EMAPS, to engage with the widest possible variety of actors, we soon had to recognize that we had little clue who these actors were or what they were concerned about. Not because of lack of candidates, to be sure, but because of their proliferation. International negotiators seemed an obvious target, but what about NGOs, local administrators, companies, climate scientists, activists, indigenous communities? What about the non-human actors involved: forests, rivers, shores, hurricanes, species threatened by extinction? To make things worse, none of these groups have clear-cut borders or evident spokespersons. Which of their members should we elect as representatives?

Had we had a clear view of how the adaptation debate was structured, we could have sampled its actors or reach for the most relevant ones. But the fluidity of the adaptation debate offered no clear landmarks for navigation. We were trapped in a vicious circle: since we had no informants, we could not improve our understanding of the controversy and, since we had only a vague appreciation of the debate, we did not know whom to engage with. We were lost because isolated, and isolated because lost.

As in all bootstrapping dilemmas, the solution comes from iteration. We can't design good maps from scratch nor summon large publics out of thin air, but we can design bad maps and then improve them, engage with small audiences and then extend them. And this is precisely what we did. We started by getting in touch with other research projects on climate adaptation (in particular weadapt.org) and asking them how we could help. At first they couldn't really tell because they had no clue of what our methods could deliver for them. So they asked imprecise questions and we returned them bad results. Slowly, mistake by mistake, the collaboration improved: they started to understand us and we started to understand them. More importantly, they put us in touch with other actors of the debate (negotiators, activists, climate scientists…) helping us start a new and larger cycle of consultation. By the end of the project, we had produced a decent set of diagrams of the adaptation debate (www.climaps.eu and Venturini *et al.* 2014) and compiled an address book spanning across a variety of disciplines and societal sectors.

Turning a vicious circle into a virtuous spiral, however, required a fundamental change in our research practices. It made little sense to organize in a classic protocol line-up where research question, data collection, analysis, visualization, and dissemination follow neatly after one another. This type of organization was just too linear and time-consuming. Had we followed it, we would have discovered at the moment of dissemination that our research questions were irrelevant for the controversy's actors and that our informants represented only a tiny minority of debate's protagonists. What we needed instead was an approach allowing us to iteratively try, fail and improve our research intervention. And this is where, learning from the experience of the Summer and Winter School of the Digital Methods Initiative in Amsterdam (cf. Rogers, 2013), we turned to the iterative and intensive format of data-sprint.

The politics of interdisciplinarity

The EMAPS example illustrates how data-sprints entail a very specific approach to scientific research and its political contribution. Traditionally, social sciences have taken two opposite but equally valuable political

stances. On the one hand, since Comte at least, researchers have supported the work of economic and administrative institutions providing them with information to uphold the organization of collective life. On the other hand, since Marx at least, other researchers have exposed the functioning of institutions providing their opponents information to contest them. Though in opposite direction, both traditions assume that the structures of collective life are given and that the aim of social sciences is to strengthen or weaken them.

This assumption is reasonable in times of social stability, but it is unworkable in situations where collective institutions are 'under construction'. Public controversies, such at the one on climate change adaptation, are a classic example of such situations (Callon *et al.* 2009). In these situations, the problem is not to support or denounce previous equilibriums, but to deal with their evaporation. In controversies, it is idle to argue about the fairness of earlier conventions, since it is precisely their breakdown that creates the dispute. What matters instead is to help social actors to work out a new cohabitation. If possible, one that is more durable and inclusive.

This is precisely the objective of 'controversy mapping' (Venturini, 2010 and 2012) an original research method developed within the tradition of Actor Network Theory (Latour, 2005). Controversy mapping (CM) is interdisciplinary by construction. Any researcher aiming for political relevance ought to reach beyond her disciplinary boundaries, but in CM this obligation becomes extremely important. For scholars practicing functionalist or critical research, it is not hard to identify the actors to engage with: they coincide either with the formal members of the investigated institutions or with their self-appointed opponents. Such leisure is not available for controversy mappers, as public debates arise precisely when the official actors (the experts, if you wish) fail to contain their disagreements. In the words of Walter Lippmann:

> Government consists in a body of officials, some elected, some appointed, who handle professionally, and in the first instance, problems which come to the public opinion spasmodically and on appeal. Where the parties directly responsible do not work out an adjustment, public official intervene. When official fail, public opinion is brought to bear on the issue (Lippmann, 1927, p. 63)

But if anyone who is concerned by the consequences of a controversial situation (as in the famous definition of John Dewey, 1946) should be considered a legitimate actor of that situation, then aren't controversy mappers forced to engage with a monstrous multitude and variety of actors? Yes, they are – and it is precisely to handle such extreme interdisciplinarity that the format of data-sprint has been introduced.

From our perspective, interdisciplinarity is not a value in itself. When things are stable enough, when uncertainty is limited and disagreement confined, disciplinary boundaries can have great virtues. They allow us to rely on previous paradigm and to take advance faster and more surely. Yet, social researchers cannot limit their intervention to such convenient circumstances. They political responsibility does not stop at the frontiers of existing institutions, but extend crucially to the moments of radical transformation. And these are also the situation where the contribution of social researchers is more needed, but also more difficult. Data-sprints are a modest but pragmatic suggestion to handle such difficulty.

References

Aykut, S. C., & Dahan, A. (2015). *Gouverner le climat ?* Paris: Presses de Sciences Po.

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world : an essay on technical democracy*. Cambridge Mass.: MIT Press.

Dewey, J. (1946). *The public and its problems: an essay in political inquiry*. Chicago: Gateway Books. Retrieved from http://books.google.com/books?id=IMkLAQAAIAAJ&pgis=1

Heede, R. (2014). Tracing anthropogenic carbon dioxide and methane emissions to fossil fuel and cement producers, 1854–2010. *Climatic Change*, *122*(1-2), 229–241. doi:10.1007/s10584-013-0986-y

Latour, B. (2005). *Reassembling the Social*. Oxford: Oxford University Press.

Lippmann, W. (1927). *The Phantom Public*. New York: The Macmillan Company.

Rogers, R. (2013). *Digital Methods*. Cambridge Mass.: MIT Press.

Raymond, E. S. (2001). *The Cathedral and the Bazaar*. Sebastopol, Ca.: O'Reilly Media.

Venturini, T. (2010). Diving in magma: how to explore controversies with actor-network theory. *Public Understanding of Science*, *19*(3), 258–273. doi:10.1177/0963662509102694

Venturini, T. (2012). Building on faults: how to represent controversies with digital methods. *Public Understanding of Science*, *21*(7), 796 – 812. doi:10.1177/0963662510387558

Venturini, T., Meunier, A., Munk, A. K., Borra, E. K., Rieder, B., Mauri, M., … Laniado, D. (2014). Climaps by Emaps in 2 Pages (A Summary for Policy Makers and Busy People in general). *Social Science Research Network*, (ID 2532946).