Building Research Infrastructures to Study Digital Technology and Politics: Lessons from Switzerland*

Fabrizio Gilardi Lucien Baumgartner Clau Dermont
Karsten Donnay Theresa Gessler Maël Kubli
Lucas Leemann Stefan Müller

July 16, 2020

Abstract

The effects of digital technology on political processes are an important phenomenon that, due to several structural problems, remains poorly understood. A key issue is the lack of adequate research infrastructures, or the lack of access. We first discuss the challenges many social scientists face and then present the infrastructure we built in Switzerland to overcome them, using COVID-19 as an example. We conclude by discussing seven lessons we learned: automatization is key; avoid data hoarding; outsource some parts of the infrastructure, but not others; focus on substantive questions; share data in the context of collaborations; engage in targeted public outreach; collaboration beats competition. We hope that our experience will be helpful to other researchers pursuing similar goals.

*We gratefully acknowledge financial support by the Swiss National Science Foundation, Grant nr. 10DL11_183120. We thank Alix d'Agostino, Hannah Stenzler, and Rocco Leonardi for research assistance, and the Science IT team of the University of Zurich, and in particular Pim Witlox, for technical support.

1 Introduction

Digital technology affects politics in many ways. The role of social media in elections, especially in connection with their potential to spread disinformation, has been one of the most visible aspects of the phenomenon. It is also one of the most researched in political science and political communication (e.g., Guess et al., 2019; Jungherr et al., 2020). However, digital technology also affects how public administration works ("e-government"), and more generally how the state interacts with its citizens (and potentially surveils them). Moreover, digital tools and platforms promise to facilitate new forms of participation and citizen involvement in decision-making processes ("civic tech").

The connections between digital technology and politics are complex, multi-faceted, and, despite a surge of high-quality research, not as well understood as they should be. The research community lacks clear answers to many questions that are highly salient to the public and decision-makers alike: what is the prevalence of disinformation on different platforms and countries? How do online political ads affect behavior and are they similar to offline ads? How can we strike a balance between data protection and the transparency of digital platforms? How can digital technology improve political participation?

One reason why answering these questions is difficult, we argue, is the existence of several structural challenges. We argue that the root problem behind such challenges is the lack of adequate research infrastructures, or the lack of access to them. We first outline the nature of these challenges and then present the infrastructure we built to overcome them in the Swiss context, which we illustrate using the example of the public debate on COVID-19. We conclude by offering recommendations for other scholars interested in replicating our efforts in other contexts.

2 Challenges of studying digital technology and politics

The first challenge is data access. Many data that researchers need to answer questions on digital technology and politics are hard to obtain, for several reasons. First, skills required to collect online data are different from what we traditionally train our students in (Salganik, 2017). Several initiatives, such as the Summer Institutes in Computational Social Science,[1] have helped social scientists close the skills-needs gap. With new graduate programs and more methods courses geared towards computational social science, many junior scholars are now trained in many of these essential skills. But even with the required skills, data access remains problematic. Researchers are largely dependent on the goodwill of digital platforms. Some, like Wikipedia, provide APIs that allow for extensive data access. Others, like Twitter, provide APIs with significant restrictions regarding the amount of content that can be accessed, and over which time period (Steinert-Threlkeld, 2018). Still others, like Facebook, have largely locked down data access. This state of affairs was described as an "APIcalypse" preventing independent, critical research on digital platforms (Bruns, 2019). Today, access to the most valuable data remains exceedingly difficult without significant resources or collaborations, effectively limiting many kinds of analyses to a select few. Initiatives such as Social Science One (King and Persily, 2019) have worked hard to provide transparent processes to gain access to Facebook data, but Social Science One "is not a one-size-fits-all model, nor is it intended to be" (Levi and Rajala, 2020, 1). Moreover, all existing efforts "are both novel and experimental. Evaluation of which is best suited for what type of data and circumstances is still in the future" (Levi and Rajala, 2020, 2).

The second challenge is data permanence. Typically, researchers collect the data they need for their projects and, when money runs out or the project is done, they stop. The data are not updated, and other researchers do not have access to them — often dictated

[1] https://compsocialscience.github.io/summer-institute/.

by the platforms' terms of use. New projects have to basically start over from scratch. This is very inefficient and significant resources are regularly wasted redoing what has already been done. A better system would be if data were collected continuously in some centralized way so that many researchers could access what they need when they need it, including for replication purposes. At the same time, hoarding vast amounts of data that nobody uses is not meaningful.

Third, data sharing is often constrained by more or less clearly defined rules. Twitter data, for example, can be shared freely only within research groups, although what counts as a research group is not entirely clear. Tweet IDs can be shared publicly, but they are not an adequate solution. These IDs allow other researchers to identify relevant posts, but they still need to be re-downloaded and pre-processed. For replication purposes, the arrangement is ineffective because tweets (and accounts) may have been deleted since the original data collection, making it impossible to reproduce the original results (Zubiaga, 2018).

The fourth challenge has to do with data protection. The European Union's General Data Protection Regulation (GDPR) has a global reach, since it affects any researcher collecting data on EU citizens. Although the GDPR includes an explicit research exception, it is poorly defined (European Data Protection Supervisor, 2020). Consequently, researchers must be mindful of the constraints set by GDPR without clear guidelines on how to navigate them.

Fifth, most research in this area is focused on one specific country, the United States (Jungherr et al., 2020, 7). Its size, language, institutional context, electoral and party system etc, are not representative of other countries. Therefore, results based on the United States might overstate certain effects as they are bound to one context, rather than controlled for in various settings. As Jungherr et al. put it, "Any uncritical generalizations on the role of digital media in politics based on cases and findings from the United States is obviously deeply naive" (2020, 7). Relatedly, research on the US case does not have to worry too much about languages. In other contexts, however, multiple

languages are relevant and constitute a challenge, despite the increasingly high quality of automatic translation and progress in NLP approaches for languages other than English (e.g., de Vries et al., 2018).

3 A Research Infrastructure to Study Digital Technology and Politics

In this section, we describe the infrastructure we built at the Digital Democracy Lab,[2] using a relatively simple COVID-19 analysis as an example. In the next section, we then discuss which broader lessons can be drawn from our experience.

Starting with our example, Figure 1 shows the salience of COVID-19 in Switzerland from January 2020 until June 2020, with a focus on traditional and social media. The analysis includes about 5 million Tweets for 181,000 users, 16,000 Facebook posts by political actors published on 156 public pages, and one million articles published in 84 newspapers. These documents are multilingual, including Switzerland's three largest official languages (German, French, and Italian). Across the three platforms, we see the striking extent to which COVID-19 has dominated public attention. The Swiss debate on COVID-19 started after the first cases were detected in Europe, and achieved a high degree of salience when the Swiss federal government enacted the first measures against the spread of the virus. Salience reached a peak when Switzerland declared the state of emergency. Then, the topic's salience gradually decreased, with spikes when the government announced new rules. Interestingly, attention to COVID-19 has been higher in newspapers (with peaks of about 70% of all articles) than on social media.

This analysis illustrates the basic workflow that is the backbone of more sophisticated studies. It requires overcoming several of the challenges discussed in Section 2, and in particular data access and permanence.

Thanks to the infrastructure we had already built, we have been able to collect and

[2] https://digdemlab.io/.

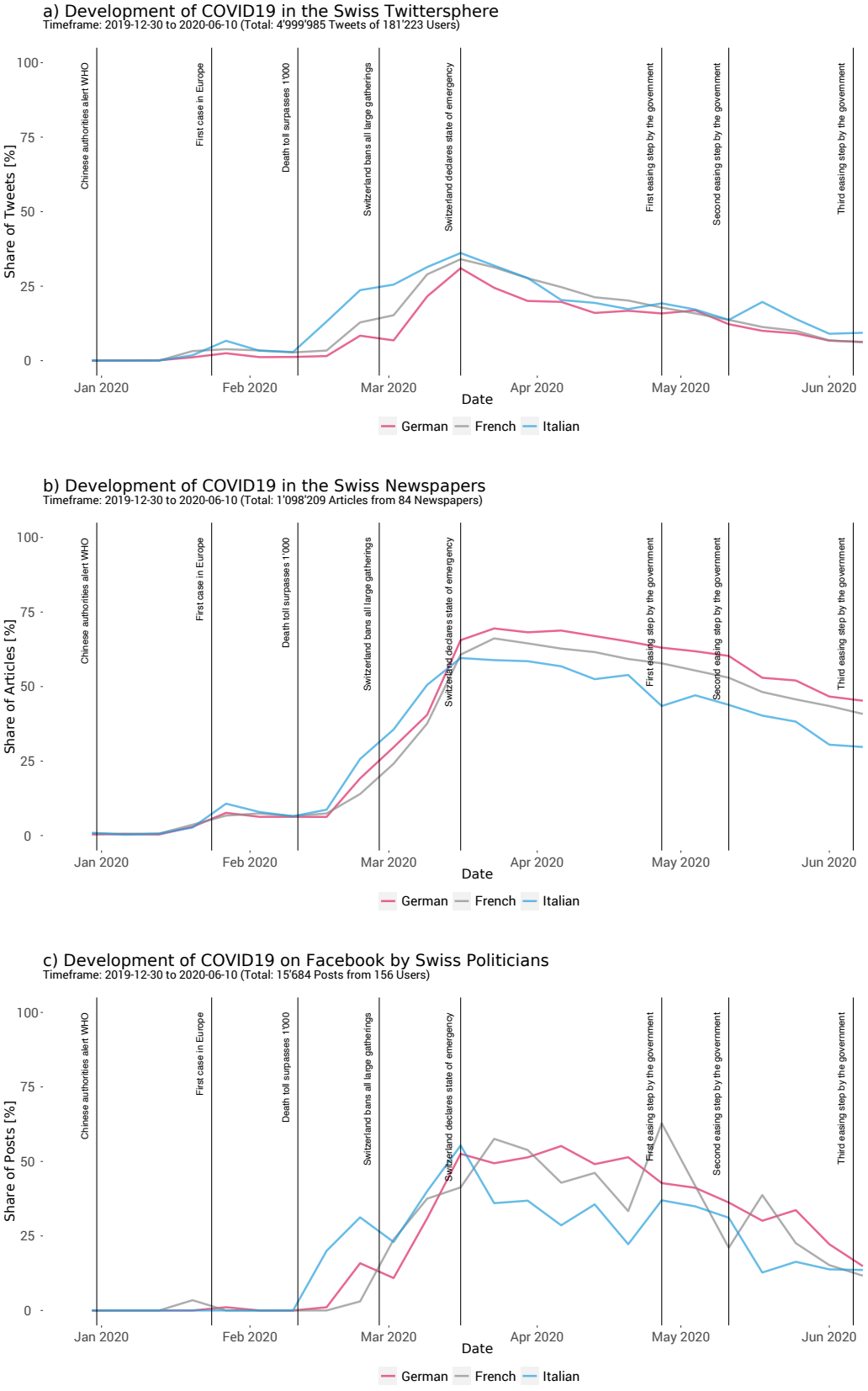

Figure 1: *Salience of COVID-19 in traditional and social media in Switzerland.*

analyze new data very efficiently. The infrastructure is shown in Figure 2. It has four components: data collection, data processing, data storage, and data analysis. Specifically, the first and second components consists of servers and scripts to carry out data collection and processing tasks which, importantly, can be scheduled (e.g., download a Twitter timeline automatically once a day). The third component consists of a database in which all information is stored and checked automatically for integrity and duplicates, once a day. The database is distributed over several servers to ensure data permanence: if there is a problem with a server, the database remains fully operational, including backup capabilities. The fourth component, data analysis, runs on additional servers with GUIs for R and Python analyses that, like for data collection, can be scheduled. For example, new documents can be classified automatically using existing scripts as they are added to the database.

In our example, we needed to collect millions of tweets, hundreds of thousands of newspaper articles, and thousands of Facebook posts. Each source has its own document format and requires different data collection and processing features. Here, our infrastructure provided us with a unique advantage: while we had to adapt our scripts (e.g. Twitter timelines) we could build on the versions we had already implemented in the infrastructure we have just described.

To build our infrastructure, we engaged with the relevant stakeholders in the IT services of our institution to build scalable solutions that implement best practices, in particular regarding database construction, tasks scheduling, and network structure. We considered a commercial cloud service, but concluded that our in-house IT services have several advantages. First, the physical and institutional closeness to the service facilitates a smooth exchange of information. Second, IT services from one's own institution are often better suited than a cloud provider to help with then kinds of problems researchers typically encounter, since they are used to working with researchers, even though not necessarily with social scientists. Third, keeping the infrastructure in-house makes it easier to comply with (local) data protection rules, as IT services have experience with

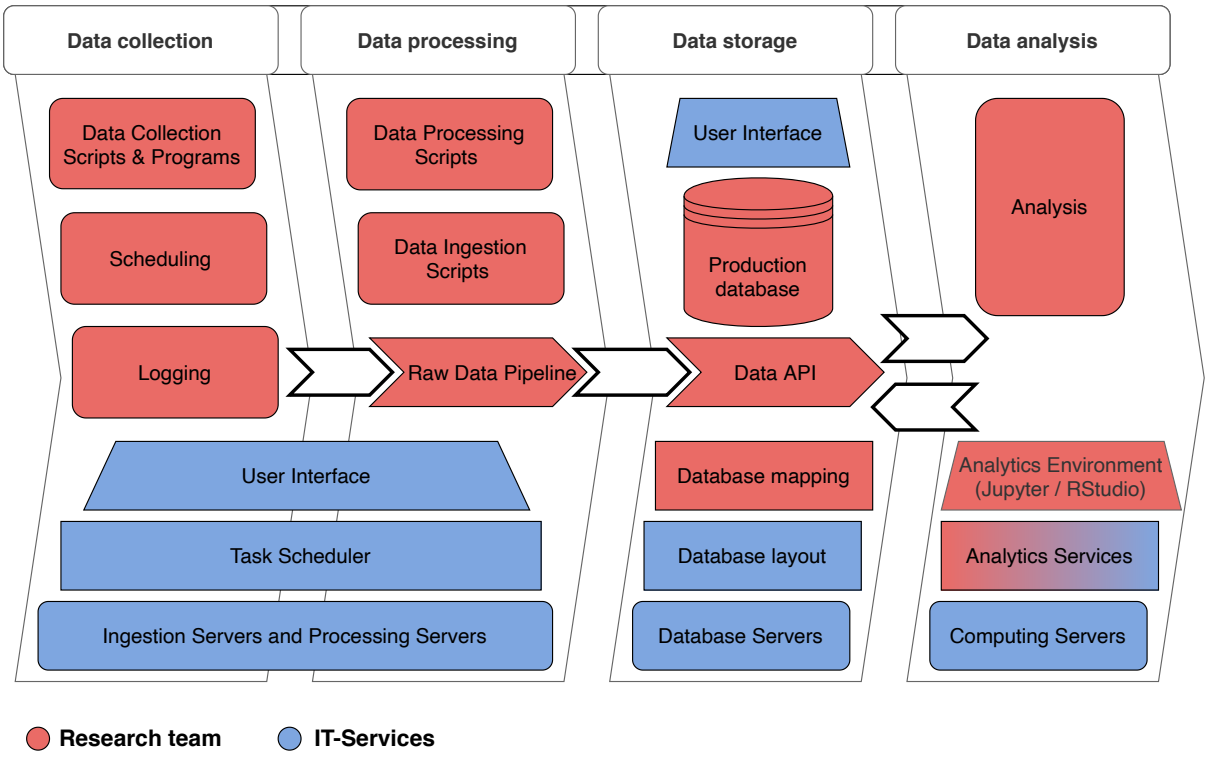

Figure 2: *Overview of the research infrastructure.*

them from other fields, such as medical research.

Relying on a professional IT service is important to ensure robust implementation of these features and guarantee maintenance, data security, and data permanence. At the same time, we keep some tasks under our direct control to ensure that we can react as quickly as possible to new needs. Therefore, one question we were confronted with was the division of labor between the IT services and the research team. Our setup is shown in Figure 2, where tasks carried out by the IT services are colored in blue, whereas those carried out by the research team are colored in red. The IT services take care of server-related back-end tasks, such as setting up the servers (including operating systems and network infrastructure), combining different machines into computing clusters, troubleshooting, and hardware maintenance and replacement. The research team does the rest. That is, we write scripts for data collection (e.g., adding new sources), processing (e.g., transforming the raw data in the desired formats), storage (e.g., how data are written into the database and implementating the search functions), and of course analysis.

This arrangement ensures a robust implementation of all features we need while keeping as much control as possible over the tasks that are most directly related to research.

Thanks to our infrastructure, we could adjust our data collection and analysis workflow very quickly to study COVID-19. Regarding data collection, we had to adjust settings on an existing server to increase system memory to process the amount of tweets, load scripts on the server, and schedule weekly data downloads. Then, we added all Twitter IDs of interest to a script using one of our functions to collect tweets from user timelines. Regarding data analysis, we could adapt classifiers we used for similar purposes, which were already fully implemented within our infrastructure. Specifically, we implemented a keyword search optimized for identifying texts related to COVID-19. It is a simple classification method that works well in this context, since COVID-19 is a topic that is discussed using a set of words which are very unique to it. However, our infrastructure can handle more sophisticated machine-learning classifiers. It would have taken us more time to implement them, but we could have done it very efficiently if needed.

In sum, the research infrastructure we have described in this section not only allows us to continuously collect large amounts of unstructured data, but it is also flexible and scalable, so that we can adjust or expand data collection and analysis routines quickly as new research needs arise.

4 Conclusion: Lessons Learned

In this paper, we have described a research infrastructure that we built to address some of the challenges inherent to the study of digital technology and politics. We now discuss the lessons that we have learned, which, we hope, can be helpful to other researchers pursuing similar goals.

First, automatization is key. Some kinds of data, such as social media, are hard to get retrospectively. Therefore, there are high payoffs in setting up an "ingestion system" that collects data continuously. Once such a system is built, new data sources can be

added efficiently and on short notice. We recommend adding new sources as soon as they appear potentially useful for current or future research. This advice, however, should be balanced against the risk of hoarding data with no clear purpose.

Second, when an efficient "ingestion system" is up and running, the temptation is to collect data just because it is easy to do so. This is not a fruitful strategy. Even though automatization reduces the marginal costs of data collection, there are still costs. Excessive collection may lead to a one-size-fits-all approach that neglects the specificities of individual data sources (e.g. different interactive features). Moreover, data collection risks becoming an end in itself. We recommend defining and regularly updating clear research areas that can help prioritize data collection. The best protection against hoarding data that nobody will ever use is to constantly be in an exchange with people that are working, or planning to work, with those data.

Third, some parts of the infrastructure can be outsourced, while others are better left under the direct control of the researchers. Most universities have a scientific IT service that can host the data, provide servers for computation, and support setting up and maintaining the ingestion system. One of the advantages of relying on the university's own service, compared to a cloud computing provider such as Amazon Web Services, is that it ensures compliance with local data protection regulations. Moreover, it is helpful to have a partner on site with whom one can establish a non-commercial relationship. In our experience, university scientific IT services are motivated to collaborate with social scientists, since it broadens their scope beyond traditional areas of operation, which might help them gain additional resources. What is less amenable to outsourcing is the interface between research and IT. We recommend that the team includes a social scientist with a strong technical background who can carry out some tasks directly (such as adding new sources to the ingestion system and making sure that everything runs smoothly) as well as communicate effectively with the scientific IT service.

Fourth, to secure funding to set up the infrastructure, it might be helpful to embed the infrastructure within a substantive project. Although it depends on the specific context,

funding for infrastructure tends to be scarcer than for substantive projects. In terms of budget, one full time position over a year might be enough, plus any additional costs that the university's scientific IT service may charge. Once the infrastructure has been established, a part-time position is in many cases sufficient to keep the system running, especially in smaller countries such as Switzerland.

Fifth, data sharing is not a trivial problem given various legal constraints. The kinds of data that such infrastructures collect are likely to be subject to restricted sharing, due to the terms of service of the platform from which they were collected and in accordance with data protection regulations. However, most of these issues arise only when the data leave the research group. Therefore, if the data cannot go to the researcher, we recommend bringing the researcher to the data. In other words, data sharing can take place in the context of joint projects with other researchers. Moreover, this strategy helps to avoid becoming a pure service provider, since data sharing is structurally tied to substantive research projects in which the core team members participate.

Sixth, the data collected with the infrastructure may lend themselves well to public outreach, as the example in Section 3 shows. Here, we recommend that researchers develop clear expectations. Like for the "data hoarding" problem, there are many topics in the political news cycle which are amenable to analysis or visualization. To make sure that this kind of work has an impact, we recommend reaching out to journalists prior to investing too much effort in a specific analysis. Impact depends on established media reporting on the analysis. Such outreach is not necessarily a key component of the infrastructure, but it is helpful to increase visibility and, potentially, funding opportunities.

Seventh, competition is generally a good thing, but establishing one infrastructure per country may be a sensible strategy, although of course it depends on the size of the country. The whole point of the infrastructure is to avoid wasting resources in duplicating data collection efforts. In this context, collaboration is more promising than competition.

The interplay between digital technology and politics is one of the most pressing chal-

lenges our societies are facing. Research on this issue faces several specific constraints. We argue that building a dedicated research infrastructure is an important step to overcome them. And we hope that our experience, discussed in this paper, will be helpful to other researchers pursuing similar goals.

References

Bruns, A. (2019). After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research. *Information, Communication & Society 22*(1), 1544–1566.

de Vries, E., M. Schoonvelde, and G. Schumacher (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis 26*(4), 417–430.

European Data Protection Supervisor (2020). A Preliminary Opinion on Data Protection and Scientific Research. pp. 1–36.

Guess, A., J. Nagler, and J. Tucker (2019). Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Science Advances 5*(1), eaau4586.

Jungherr, A., G. Rivero, and D. Gayo-Avello (2020). *Retooling Politics: How Digital Media are Shaping Democracy*. New York: Cambridge University Press.

King, G. and N. Persily (2019). A New Model for Industry–Academic Partnerships. *PS: Political Science and Politics*, 1–7.

Levi, M. and B. Rajala (2020). Alternatives to Social Science One. *PS: Political Science and Politics*, 1–2.

Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Steinert-Threlkeld, Z. C. (2018). *Twitter as Data*. Cambridge: Cambridge University Press.

Zubiaga, A. (2018). A Longitudinal Assessment of the Persistence of Twitter Datasets. *Journal of the Association for Information Science and Technology 69*(8), 974–984.