Data, In Practice: risks, pressures and ethics for data practitioners

Elevator pitch

This project examines how we can help practitioners learn about risks and good practice related to digital data collection, processing, and storage, and incorporate good data management into their day to day work. We start with practical guides and interactive tools for helping organizations build data policies and practices that are fit for purpose, but we need also to look at the drivers of good and bad practice, and tackle them through research and advocacy.

> *"if you want deep sustainable change, then this "Data, in practice" research and practical tools needs to help"*
> *- Heather Leson, Data Literacy Lead IFRC*

The problem we're solving

Digital tools have empowered nonprofits and civil society actors to collect, store, and process increasingly more—and increasingly more sensitive—data in the course of their ordinary service delivery. Alongside new opportunities, as data becomes increasingly intertwined with an organization's core programmatic activities, there are new considerations: new obligations to protect and responsibly handle client data, and new risks, as the interaction between data and societal challenges creates new potential harms for vulnerable groups.

To these programmatic risks, we can add operational and reputational ones, as governments begin to regulate sensitive data collection, transmission, and storage. These new laws are scattered across different areas of regulation, and for international organizations, may be hard to balance across jurisdictions. Many civil society organizations are simply unprepared to process these emerging requirements, and do not have the resources to manage compliance even if they could make sense of them.

At the same time, conflicting ecosystem pressures complicate the way we talk about responsible data. The conversation to date has been dominated by the push for open data at all costs, and the popularity of 'big data', so that organizations feel pressure to collect as much data as possible, 'while we're here', even without a specific purpose at the time or effective management measures in place. Meanwhile, the focus of civil society dialogue and training has been on digital security, recommending measures such as hardened file transmission systems and encryption. Yet such approaches can ignore the human and organizational factors which present real risks to data subjects. Humans make mistakes, or use insecure workarounds for

secure systems to make everyday life easier. Organizations create complex and insecure data collection workflows, and don't think about how spreadsheets full of personally identifiable information are collected, stored and transmitted. For example, very few organizations have a policy or a process in place to track data deletion.

Available resources mirror these problems. Many are preoccupied with 'big data' projects and solutions - with data exhaust from existing government projects, other large datasets, and their use and handling. Few deal with projects of the size and scope of most that we encounter at SIMLab - delivering services to fewer than 1 million people at a time, researching the behavior of communities and sub-groups, gathering monitoring data or bank account information in order to assess or deliver help. Those that exist are too complex, or in draft, and all are in English. None effectively address the critical legal issues that arise.

What we've done so far

In 2017, as part of the Good Data Collaborative, a consortium funded by Stanford's Philanthropy and Civil Society team, and including the Engine Room, the Center for Democracy and Technology, and the Future of Privacy Forum, to set out the state of the field as it stands. We are:

- Redesigning ResponsibleData.io
- Carrying out a literature review and gap analysis (published September 2017)
- Conducting a user consultation of practitioners to determine what their needs are and publishing our findings in a short report (published September 2017)
- Developing basic prototype tools based on that input (by end 2017)

Our research this summer has confirmed that, despite the scholarship being produced in this field, organizations need for more practical guidance, particularly for organizations for whom data is not a core part of their mission. By the end of the year, we will have developed machine-readable, online resources hosted by the Engine Room's ResponsibleData.io site to provide a first port of call for practitioners learning about RD for the first time.

However, our findings suggest that tools and resources alone are not enough; organizations need to institutionalize RD approaches in their risk management, human resources management, and data management procedures. For this, they often need infrastructure investment. And we must address the growing disconnect between those who acknowledge the importance of responsible data (whether rooted in respect for the rights of data subjects, or concern for their protection, or both); and those who believe that data is this century's big extractive industry, and that business models which transparently monetize data, and using large datasets and algorithms to guide decisions and generate insights are effective ways to help vulnerable people. We must acknowledge the outsize power that donors and philanthropists have to direct these conversations and point their grantees towards, or away from, good practice.

What we need to do next

We see three streams of work emerging here:

- Developing better tools and resources, machine readable or translated and geared at different levels, from beginners seeking to improve their approaches swiftly to people looking for more in-depth templates for organizational tools

- Testing different approaches to supporting improved RD practice in organizations. Our project proposal for this is below.

- Addressing the disconnects in the conversations we are having with advocacy, discussion and research. This is particularly important at donor level. Our research so far has hinted at it, but more is needed to uncover how indirect 'nudges' and incentives could be employed - or removed - to influence practice. We also need to confront differences of opinion here between powerful influencers and investors who are actively promoting data as the next big extractives industry. These conversations must be opened up and the relative positions understood, as fluently as we do policies around gender and accountability.

There are a number of actors working in the space, however, and many of the discussions we're having are converging. We should also ensure that we bring everyone together for dialogue in specific fora, to coordinate our effort.

Tools and resources

Below are some of the ideas we have for resources that could be developed. Where indicated, some of this work is already underway - the rest is not yet funded or will be done by us pro bono as we have time.

- **101-level:** online, machine-readable, and designed for people who are new to the content, the resource we will develop with the Engine Room as part of the Good Data Collaborative will allow users to click through guidance according to their needs, and get extra resources when they are ready for them. [Funded, delivery by end 2017]
- **Diagnosis:** Using an improved version of Friendly, a prototype tool SIMLab developed that helps users assemble their own question-and-answer / document builder applications without using code, create a system which helps practitioners identify where they are and where they'd like to be, and generate tools and policies to help them bridge the gap. Friendly is a riff on existing form builders, with a focus on custom logic that runs without a server.
- **Legal:** Produce a short brief on the legal exposure inherent in data management. Although we would not provide legal advice, we may consider partnering with law firms

who do pro bono work in the field, to provide additional avenues for organizations that need more specific help. More ambitious levels of this work might include mapping what we know about the law on this in different countries, and available guidance and advisors.

- **Web-based course:** Develop an email-based curriculum based on our resources to help train busy project officers - this course would be open access, meaning that it could be easily translated and picked up and customized by any organization as long as the license is maintained

Responsible Data Sprints: An adaptive experiment on drivers of organizational change

We propose to test a number of interventions which could be put in place at organizational level to radically improve visibility and understanding of responsible data handling practice.
We will be working with Feedback Labs, a consortium of over 300 organizations working to create new norms in the philanthropy, aid, and governance sectors that put regular people in the driver's seat, with experts in a supporting role during program selection, design, implementation, and evaluation. Members of the Feedback Labs network are committed to asking three simple but important questions: What do people want to make their lives better? Are we helping them get it? If not, what should we be doing differently?

From Feedback Labs' network (and if needed, other interested organizations) we would assemble 12 test organizations and separate them into 3 cohorts. Each test organization would be audited to establish a baseline of responsible data practices. Each cohort would then be supported in three consecutive, 3-month sprints to test specific approaches to building data responsibility:

- **Cohort 1: Toolkits**. Participants are provided with tools and guides and encouraged by the organization to implement them in their work.
- **Cohort 2: Organizational compliance & culture-building**. We support the organization to roll out a responsible data policy. This is combined with work with the management team to envision what a values-based approach to including RD in their work would look like in practice. RD is included in job descriptions, person specifications and employee evaluations; it is included in organizational risk registers and project management; and project monitoring.
- **Cohort 3: Individual capacity-building**. Individual staff members become 'champions', are provided with mentors and training and encouraged to change the culture of their organizations.

In three-month sprints, SIMLab and Feedback Labs will work with participant organizations to roll out the interventions. At the end of each sprint the audit is repeated, to see which

approaches are having the most success. The least successful is dropped after the first two sprints, OR approaches are rolled together or remixed and a new rollout carried out, as is determined to be most impactful and appropriate.

By the end of the project we should be able to make informed statements about which approaches have most successfully improved responsible data practices within practitioner organizations. This is not research, but adaptive experimentation. We don't expect to see robust, distinct differences between the approaches taken in different cohorts, particularly as many of the measures can lead to those in other interventions. We expect the cohorts to bleed together to some degree by the end. Rather, this is about the starting point and the most productive way to support organizations and individuals on a journey towards responsible data practices. Practical guidance—by way of checklists, apps, or other tools—will be made available via the three organizations involved, and can also be integrated into the incentive framework of GlobalGiving Rewards.

Coordination, reality checks and frank conversations

Our research during phase I of this work revealed confusion and lack of knowledge about this issue not just among practitioners, but among donors too. Some had no-one covering this issue on their staff, and referred me to potential interviewees who are only tangentially related. None provided systematic support to grantees for responsible data practice.

There is significant work ongoing in this area, though, and some is not as coordinated as it could be. The excellent Responsible Data listserv run by the Engine Room tends to be used to report out, not to coordinate up front, as competition for funds means that organizations don’t share plans ahead of proposals.

- Map people working on this, by sector, and ensure they come together at one of the existing events, for example, those run by Stanford PACS. Run policy-level side events for progress-sharing, coordination and frank discussion.
- Develop two page explainers for donors, sharing existing tools and proposing additional support for RD (not just digital security) for grantees (not just the donor themselves)
- Share research findings and provocative messaging through videos, social media, blogging and op-eds linked to relevant themes. Raise with IRIN News, Reuters AlertNet, and the Guardian Development desk among others, to support investigative journalism around the potential breaches and dangers here.
- Look ahead to identify potential future and emergent risks and issues relating to data. For example, an interesting angle would be to examine what, if anything are the implications of AI and algorithms as emerging tools for social change work on grassroots implementers? How will they and their clients be subject to them, and to what extent can or should they make use of them?

What will the project leave behind?

Up to 300 words: how is this sustainable after we've finished the work? What learning will be shared and how? Do we have a plan for a graceful exit from the beginning?

As with all of our resources, the tools we develop will be freely accessible on the web, licensed under Creative Commons (SIMLab's default is BY-NC-SA), machine-readable, downloadable and without any registration walls. Users should be able to generate their own data policies and risk assessments without divulging any personal information to us. (We may collect aggregate data on how people answer the policy-builder questions.)

The most immediate beneficiaries of our work are small and medium-sized nonprofits and civil society actors that work with sensitive data. We hope our tools, research and advocacy will build capacity for these organizations to ask better questions about data risks, and take practical steps to mitigate them.

In addition to making the tools and our development process publicly available, we intend to share them with our membership networks, including the Responsible Data Forum, Feedback Labs, CaLP, OpenGov Hub, and the CDAC Network. Because the work will necessarily involve participatory networks of partners and clients, we intend to use those networks to share back draft and final versions of our work.

Throughout, we plan to document our thinking and our development processes in a series of blog posts and other media, from staff, partners, and clients.

About SIMLab

Our work directly addresses structural inequality in the digital age, seeking to understand how communities in crisis communicate, right now, all over the world.

Our communications strategies target the hardest-to-reach, people who are often struggling with overlapping challenges, systems that don't work for them, and basic barriers to access. Putting information and access to services in people's hands using the devices and systems they already have in their pockets and homes can empower them to make their own decisions about their lives; give them access to lifesaving emergency information; give them the tools to hold institutions to account; and help them to make changes and get help so they can break them out of cycles of disadvantage and become more resilient to shocks.

We utilise 'inclusive' technology, from radio to social media, from SMS to Whatsapp, from the mobile web to community noticeboards, these platforms are accessible and easy to use; the hardware they require is relatively inexpensive to buy and maintain, with parts available locally, and manageable running costs (power and airtime, for example). They may not be digital. They

must be part of a two-way feedback loop. The important thing is to ask what people use in a place and time, and more importantly, what they trust and act upon.

We help organizations providing information and services to communities in crisis to develop interaction strategies and tactics that meet the most vulnerable where they are.

Social Impact Lab (SIMLab) began life as a non-profit with a mission to lower barriers to social change through mobile. SIMLab was the institutional home of the FrontlineSMS project, a suite of software that helps organizations build services with text messages. With more than 250,000 downloads in 199 territories, FrontlineSMS helps tens of thousands of organizations reach tens of millions of people using SMS, customizing and owning their own content, engagement, and relationships in sectors ranging from health, to financial services, to conservation. **SIMLab's impact has been recognized by the NGO community at large: in 2013, the Global Journal named us the #1 Tech NGO in the world, and we received a Google Impact Award jointly with Landesa. In 2011, SIMLab was awarded the Curry Stone Design Prize.**