Data Federations: Digital Collaboration Without Data Sharing

Document - ODI Report
Updated - 15.03.2021
Author - Richard Woodall

etic lab

Data Federations: Digital Collaboration Without Data Sharing

Document - ODI Report
Updated - 15.03.2021
Author - Richard Woodall

Glossary

Data Federations (DF): A new model for the organisation of digital collaborations, based on the use of distributed privacy-preserving technologies such as Federated Learning and Distributed Data Mining. These technologies allow partners to contribute their data to a common project without requiring that the data is shared, exposed or held in a central location. By taking advantage of the possibilities afforded by FL, DDM etc., collaborators can work together to create shared value from their data whilst mitigating some of the barriers and risks that such collaborations inevitably incur.

In principle, DFs offer a way for actors who would otherwise be unable (for reasons shortly to be stated) to participate in digital collaboration to access the transformative benefits this kind of work can offer, within a safe, ethical and compliant framework. The purpose of this project is to understand which organisational structures and protocols are best suited to help organisations realise these possibilities.

Federated Learning (FL): a new machine learning technique which allows predictive models to be trained on separate datasets without any of the data being shared or exposed. This enables us to create algorithms from large, distributed datasets without infringing the privacy or security of any contributed data.

Distributed Data Mining (DDM):a method for generating insights from distributed datasets, without requiring that the data is collected in a central location or held in a standard format. Provided we understand enough about the schemas used by the different datasets, is possible to design DDM modules which can address queries to the distributed data as if it were in a centralised, standardised format.

Synthetic Data: a mechanism through which real data can be translated into an artificial dataset which preserves significant patterns from the original data without reproducing any sensitive or personally identifiable details. This can then be used for analysis or training whilst protecting the privacy of the original set.

Perturbation: a set of techniques which involve introducing elements of random of “noise” into the data in order to obfuscate sensitive details. This can be calibrated to strike a balance between security and usability.

Introduction

Creating Value in a Digital Economy

In a digital economy, value is not derived primarily from the ownership of data assets, but the capacity to do things with them. An organisation may technically own vast quantities of high quality data, but if all it does is sit inert on a hard drive - perhaps making a bashful appearance in annual reports - then it is effectively a dead asset. Data only comes into its own as a generator of value when it is combined with other information sources and used to create useful insights or predictive models, which in turn provide strategic intelligence to help organisations accomplish their goals or realise their principles in the world.

The general benefits of strategic data use are increasingly well understood. However, many organisations are still unable to engage with these possibilities, even if they would like to. Fundamentally, this is because getting the most out of data means working at scale. Technologies such as data analytics or machine learning need large quantities of diverse data to provide their best returns, not to mention the resources and expertise to be properly deployed. For many organisations, these requirements are beyond their individual means. Therefore, the only way for them to access the benefits of data will be through collaboration with others.

By working together, a group of partners can assemble the resources to attempt a digital project that none of them would be able to attempt on their own. Just as importantly, they will also have a better chance of marshalling a social consensus powerful enough to act on what their project produces. Clever insights and predictive models will mean nothing if they are not backed by a coalition of actors committed to using them in a purposeful and effective manner. Collaboration is not just about putting together the tools to address a technical challenge, but building the social formation that will use data to further its mission.

However, working together brings its own set of challenges. During our previous work in the UK's Access to Justice sector (documented in our report, *Digital Technologies in the Access to Justice Sector: A Strategic Overview*), we developed a firsthand understanding of the barriers that can stand in the way of digital collaboration. Whilst these are grounded in our experience of the charitable legal advice sector, we believe that many of the problems identified are generalisable to any circumstance where organisations with limited means attempt to work collectively with their data. They include:

- Rules around the exposure of personal data, which in many cases makes the pooling or sharing of data legally impossible. Even where sharing is not legally prohibited, organisations may still feel uncomfortable sharing information if they feel there is ambiguity regarding its sensitivity.

- Unwillingness to share proprietary information regarding business operations with organisations who may be competitors or have regulatory oversight

- Lack of time, resources and expertise to dedicate to digital projects, particularly involving cleaning and standardising large-scale datasets

- Lack of experience of successful digital projects

These are fundamental problems without easy or obvious solutions. Nonetheless, they must be addressed if we are to create a more equitable and socially useful digital economy.

The Case For Digital Collaboration

As with every other aspect of our society, our current digital economy is extremely unequal. There are many different ways of understanding and describing this inequality, but we have been struck in particular by the concept of "vectoralism", described by the philosopher McKenzie Wark.

According to Wark, the "vectoralists" constitute a new kind of ruling class, whose power is not based on the possession of land or capital, but ownership of the systems through which data is collected, processed and - most importantly - made meaningful. This allows them to construct powerful information asymmetries, through which they are expand their spheres of influence.

Take Google as an example. Google provides internet users with a range of wonderful services, many of them for free. Through its search function, it enables billions of people to access otherwise inaccessible pieces of knowledge. However, as Wark points out, while every individual gets to see the "little piece of information" they were looking for, Google gets all of the information generated by those billions of searches in the aggregate.

This is a resource of almost unimaginable power, which Google has used not only to dominate the market for search, email, digital advertising and so on, but also to expand its operations into new markets. The digital insights available to Google and the like often enable them to outflank traditional market incumbents, offering their services at greater convenience and lower (or no) cost, whilst its accumulated capital always leaves the option of buying out the competition. Google's recent acquisition of health-tracking device company Fitbit, for instance, is likely a prelude to an aggressive move into health insurance provision.

We don't want to live in a world where a handful of unaccountable corporations get to administrate the necessities of our existence. We're encouraged, therefore, by the recent regulatory backlash against the power of big tech in the USA, EU and elsewhere. However, regulation will only get us so far. In order to build a fairer and more open digital economy, it will also be necessary to redistribute the capacity to collate, interpret and act upon data at scale, such that our societies are able to enjoy the benefits these activities can provide without having to rely on the intermediation of giant tech companies.

This means we need to build up the capacity of a different set of actors - unions, co-ops, civil society organisations, professional services, businesses - to do things with their data. It also means inventing new kinds of institution to provide individuals with ways to pool their data (and their digital rights) for the common good. There is currently a wealth of exciting work going on in this area, from the ODI's work on Data Trusts to a range of recent proposals for how the state might support a fairer and more democratic digital economy. Our hope is that our work on DFs might provide a unique contribution to this discussion by drawing attention to a set of technologies and organisational methods which have not hitherto received significant attention.

What is a Data Federation?

What are the Possibilities of Federated Collaboration?

The concept of Data Federations is rooted in the recognition that for many organisations, digital collaboration is essential in principle and impossible in practice. During our time in the Access to Justice sector, we determined that if we could find some way of addressing this paradox, we would be making a real contribution to the development of a more accessible digital economy. Our search led us to consider a range of **privacy-preserving or distributed digital technologies** - some very new, some more established - which allow users to run analytics or machine learning projects without requiring data to be shared, exposed, centralised or standardised.

A key example of such a technology is federated learning - a relatively new technique which allows machine learning algorithms to be trained on distributed datasets without collecting the data in a central location. Google has used FL to train its text prediction software for mobile phones; it is also deploying it as part of its efforts to develop "privacy-preserving" methods for serving targeted ads. Applied in a context like the Access to Justice sector, FL could allow organisations to enter into collaborations which would otherwise have been impossible without requiring them to cross ethical or legal red lines or engage in technical work beyond their means.

Alongside FL, there are a range of other technologies (listed in the glossary) which provide similar benefits, allowing insights or models to be developed from data whilst protecting its privacy, reducing the need for standardisation and centralisation, or both. Each of these options offers a set of possibilities which, on their own or in combination, might be more or less applicable to the needs of a particular situation. For now, though, we'd like to focus on what they enable in general.

The true value of technologies like FL is not that they preserve privacy or reduce the need for data standardisation, but that they offer a platform for a different kind of digital collaboration. In other words, they ought to be understood not just as "defensive" measures, but also in terms of the new possibilities they might afford. By this, we mean the following:

- By providing tools which allow organisations to proactively engage with issues of privacy and non-standard data, these technologies can support collaborations which would otherwise have been impossible even to attempt.

- In this way, they open the possibility of collaboration to a range of organisations who would hitherto have been incapable of participating in this kind of work, either because of a lack of resources or because the data they work with is too sensitive to share.

- Since the costs and risks of entry are lower, these collaborations can potentially draw from a broader range of partners, allowing for the assemblage of more diverse datasets and the possibility of experimenting with more democratic and participatory forms of governance.

- By reducing the need for all participants to commit to collecting and storing their data in the same way, or obliging them to reveal their data to their partners, a federated approach can accommodate different working practices, values and assumptions around the use and value of data within the context of a shared project. For instance, one actor may believe a particular data point to be a proprietary secret whilst another does not. Whereas that would otherwise be an intractable barrier, now it can be addressed within the design of the collaboration.

- In the first instance, the product of a Data Federation will be a shared tool or set of insights, as opposed to a shared dataset. Whilst this poses its own set of ethical and governance challenges, there are substantially fewer possible externalities and dependencies in the former case as opposed to the latter. This could mean that DFs represent a less onerous governance burden for participants, although this remains to be demonstrated in practice.

- Compared to projects which require that a range of non-standard data is harmonised and collected in a single location, DFs offer the possibility of creating useful insights much earlier into the project lifecycle. This could facilitate a more attractive incentive structure for participants.

- By allowing collaborators to create shared value whilst managing or reducing costs and risks, DFs can provide an environment in which partners can build shared understandings and collaborative relationships, laying the groundwork for more ambitious projects in future. From a technical standpoint, the DF model is highly scalable, as technologies like DDM and FL make the process of adding new datasets to the federation relatively straightforward.

How Should a Data Federation be Organised?

The possibilities we have just described are not inherent in the technologies themselves. If they are to be realised in a responsible, ethical and effective manner, then the proper social structure for organising their implementation must also be designed. We have developed a set of general principles laying out the structural prerequisites and basic competencies of an effective Data Federation, as well as a step-by-step model for how to construct one (a summarised version of this is included as an appendix - more detail can be found on our website).

In brief, a fully-formed DF is a group of organisations with a common purpose and shared goals, which they have successfully translated into a project which uses their collective data resources to create shared value without requiring that they are held in a central, standardised dataset.

The DF will be governed according to a contractual agreement collectively drafted and ratified by all participants. This will define the rights and responsibilities of all partners, identify and address relevant liabilities and ethical concerns, and describe procedures for the fair distribution of value and good management of the products of the collaboration.

The DF will also produce a technical design, which will be built into a bespoke digital tool created to achieve their specific goals within relevant constraints (i.e., preserving data privacy, working with non-standard datasets etc). Owning and managing this tool and the insights it produces will be one of the DFs primary tasks.

The next step was to test these ideas in practice - which was the main purpose of our project with the ODI Stimulus Fund.

Research Questions

In the most general sense, the purpose of our project has been to road test the approach laid out above with real-world collaborators in the charitable sector. Our specific research questions were as follows:

- Would we be able to communicate the value of our approach to potential partners such that they would prepared to discuss and scope a project with us?
- Would the DF approach allow partners to consider and scope projects which otherwise would have been impracticable or impossible?
- Could we help our partners advance their goals by using the DF model to generate useful insights from their collectively contributed data?

We were also interested in what we might be able to learn about the design and construction of successful DFs. Specifically, it was our assumption that one of the main consequences of offering tools which may be used to mitigate common issues such as privacy or non-standard data would be to reveal further barriers to collaboration, possibly around the alignment of incentives, the difficulty of generating common understanding and so on. A further research question was, therefore:

- Assuming the DF approach is capable of mitigating some of the most prominent barriers to collaboration, what other barriers remain, and how might these be addressed within the design of a federation?

Our plan was to begin this process from the ground-up, by seeking to source a group of potential collaborators from amongst our contacts in the Access to Justice Sector. We were conscious that this was an extremely ambitious undertaking, with no guarantee that we would even succeed in convincing someone to give us a hearing, let alone create an actually functioning DF. Nevertheless, we reasoned that by setting ourselves the highest available goal, we would provide ourselves with the best opportunity to learn from our experience, even if in the end we were unable to fulfil our hopes entirely.

The project unfolded in three distinct stages, which will be described below.

Stage #1: Finding Collaborators

As it happened, we ended up launching our project in the midst of the Covid-19 pandemic. This proved to be something of a double-edged sword. On the one hand, it meant that the charities we were in touch with suddenly had far less free time and resources to dedicate to anything outside of the maintenance of their daily activities, much less an experimental project such as ours. Organisations who we'd previously had positive contact with were now unable to engage with our project to the extent they would have hoped. On the other, the multi-dimensional social crisis sparked by the pandemic meant that collaboration became an existential necessity throughout the Access to Justice sector and the charitable world more broadly.

A sector-wide conversation developed around how organisations might best share resources and knowledge during these trying times. Thanks to our existing relationships within the sector, we were able to be a part of these conversations, which gave us an opportunity to present the DF model to decision-makers in the sector as a way for them to address the vast challenge which currently faced them. Given these circumstances, we formulated a strategy which we hoped would give us the best chance of securing collaborators to work with whilst also providing useful insights into the process of building a DF from the ground-up.

Our strategy was as follows:

- Use our existing contacts within the sector to identify organisations or individuals with a particular interest in digital collaboration
- Attend sector-wide events to better understand the priorities and needs of the sector, as well as to sector-wide events to get our message out and make our pitch to potential collaborators
- Build relationships with key decision-makers who could build our credibility and help convene groups of potential partners

This approach required a great deal of patience and perseverance, but also gave us the opportunity to put our ideas in front of a range of different audiences. In particular, it helped us get a sense of which aspects of our pitch were connecting with people, and which required better explanation. In general, we learned to keep the technical element of our proposal as simple as possible, essentially condensing it to the proposition that “digital collaboration without data sharing” is possible, and to focus instead on inviting our interlocutors to imagine what would be possible assuming that issues like privacy, security and non-standard data were no longer intractable barriers.

Eventually, we were able to engage with two organisations who were interested in taking a closer look at our proposition. Before we describe these projects in more detail, a few reflections on the experience of seeking collaborators during these extraordinary times.

Learning #1: Whilst the circumstances we faced were in many ways extremely difficult, we also began with several advantages. We had long-term relationships with key actors in the sector, and the credibility that comes from having recently published a research report on the use of digital technologies to support the provision of legal advice. Without this, our job would have been considerably more difficult - which suggests that sector-specific knowledge and relationship building are key assets for an organisation looking to be in the business of facilitating digital collaborations.

Learning #2: On a similar note, insiders to a sector will always have a considerable head start when attempting to establish digital collaborations. In many ways, this is as it should be, since they will be in the best position to understand the purpose and feasibility of such projects. This suggests that raising the general awareness of the possibilities of digital collaboration within a given sector would be a worthwhile project, since it would incentivise organisations to engage with these possibilities on their own terms, rather than having them pitched to them by outsiders.

Stage #2: Engagement

As mentioned above, we were able to engage with two organisations in order to discuss our ideas and explore the possibilities they might have for supporting their goals. These were the Social Economy Data Lab (SEDL) and Access Social Care. SEDL represents a group of social investment organisations (i.e., organisations who specialise in provided finance to charitable initiatives), and has spearheaded a long-standing digital collaboration between its members. Access Social Care works to build links between the Access to Justice and social care sectors, with the goal of ensuring that carers and care recipients are able to fully exercise their legal rights.

Both organisations were committed to the core aim of using shared data to develop insights which would enhance their sector's ability to carry out its mission. In the case of SEDL, this meant sharing intelligence regarding trends and patterns in social lending in order to promote lending strategies which provide the best support to impactful charitable initiatives. For Access, this meant consolidating information regarding people's legal needs and the extent to which they are currently being met, with the goal of highlighting and redressing gaps in current provision.

Before describing our experience with these partners in further detail, it's worth taking a moment to highlight what these two organisations have in common, and what we might learn from these observations:

Learning #1: Both organisations were in charge of extant data collaborations, which had already achieved a certain degree of success. They had each assembled a group of collaborators committed to the principle of digital collaboration, and had gone some way towards designing and executing a shared project. This meant that our proposals spoke directly to their practical experiences, rather than a set of abstract propositions about the value of data, working together etc.

Learning #2: Both collaborations were driven and sustained primarily by the efforts of a highly talented and committed team at a convening organisation. Ultimately it was their initiative, relationship-building efforts and motivating energy which got the collaboration off the ground and provided its sustaining momentum. Otherwise, the collaborations demonstrated little in the way of formal institutional structure.

Learning #3: Whilst each project was unique in its own way, both had found that the issue of data standardisation was proving to be a considerable problem. The work involved in standardisation was simply too specialised and onerous to be done by collaborating charities, most of whom lacked the time and resources to commit to the task. This meant that the convening organisation had to take on the burden of doing this work themselves, all whilst being subject to the same time and resource pressures as their partners. That they have been able to achieve as much as they have is testament to their hard work and ingenuity, but both organisations had realised that the issue of standardisation was a major barrier towards scaling and developing the collaboration going forward

In both cases, the appeal of our approach was that it might provide a way for these groups of collaborators to generate some shared value without first having to go through the work of standardising everybody's data and gathering it in a shared dataset. In an initial meeting with both organisations, we were able to explain our ideas with sufficient clarity that our interlocutors were interested in hearing more, which supports our basic premise that the principle of "collaboration without sharing" has something to offer people who are interested in collaborative data projects.

Whilst we had a series of productive conversations with both parties, our relationship with Access ultimately did not progress beyond initial discussions and planning. Though our experience with Access did inform the learnings we will present later in this document, the following passages will focus primarily on our work with SEDL, with whom we were able to complete the Scoping stage of the Collaboration-Building process.

Stage #3: Scoping

The goal of SEDL is to encourage the social investment sector to make better use of data to inform investment strategy. To this end, they have convened a group of social investment organisations who have committed in principle to this overall aim. The data in question primarily consists of the loan books of each partner organisation, which contain information relating to the deals they have made with a range of different charitable initiatives, including details on the recipient (name, location etc) and well as the deal itself (% of grant/loan/equity, duration, interest rate etc). Naturally, each organisation has its own procedures for collecting this information, and uses different software to store it in various formats.

In order to facilitate sharing between these organisations, SEDL has designed a comprehensive data specification which provides a standard format for recording the key information required to generate systemwide insights into the social investment ecosystem. In addition, they have built a dashboard which can translate this data into useful charts and visualisation. However, they have been held back from building on these impressive achievements by high levels of time, effort and expertise required to translate the data of their partner organisations into the specification.

We believed that the DF approach had the potential to allow SEDL and their partners to generate useful insights from their data without subjecting themselves to an implausible amount of digital labour. Our basic proposition was to build a DDM module which would be able to identify and accommodate differences in the way that each organisation was recording a functionally equivalent piece of data (i.e, time, location or interest rate). This would enable us to put specific queries to the datasets of all partners without requiring that they were first translated into a standard format - effectively, bringing the specification to the data rather than the data to the specification. In this way, it would also be possible for partners to contribute their data to the project without exposing it to the other participants, which might ease concerns around the sharing of proprietary business information.

Having established the feasibility of this idea in principle, our task was to explain the concept to SEDL and their partners, and then invite them to help us define a set of questions which they would be interested in putting to their data. To this end, we arranged a workshop with a core group of social investment organisations in which we introduced them to our ideas, and invited them to brainstorm the kinds of insights which they would like to be able to derive from a collaborative data project. This left us with a range of general topic areas subdivided into specific questions, providing the basis for a set of queries that could plausibly be addressed to the DDM module.

The next job was to compile a register of available data resources and the terms upon which the partners would be prepared to contribute them to the project. To accomplish this we designed a questionnaire which encouraged our correspondents to list their available data assets and describe any pertinent constraints. As it transpired, the majority of the partners did not feel equipped to complete this document without support, so we arranged a series of one-on-one conversations where we would be able to guide them through it, and discuss the issues at hand in greater detail.

These conversations were extremely educative, providing the source of many of the learnings we will be sharing in the next section. Overall, we found that all partners were happy to contribute their data in principle, but were generally a) not clear on the value that our project would provide to them, and b) requested that Etic Lab draft a bespoke data sharing agreement describing the terms of access to their data.

This latter request was of course perfectly understandable, but also at odds with the collaborative model we were trying to deploy. In the event that Etic Lab had entered into individual agreements with each participating organisation, we would have effectively built another layer of separation between the collaborators - the opposite of what we were trying to achieve. The only way to move forward would be to draft and ratify a collective agreement between all parties as per Stage 2, but we ultimately found it difficult to motivate partners to engage with this suggestion. We scoped a range of alternative strategies for moving the project forward, but time constraints prevented us from pursuing them.

To summarise, by the conclusion of the project we had achieved the following:

- Created a design for a project based on the interests and priorities of the SEDL partners, including a technical spec for a DDM module based on their specification

- Gained practical experience of sharing and implementing the DF concept in a real-world setting, developing insights into how best to communicate and build consent for these ideas

- Developed an understanding of the broader social and organisational barriers to collaboration which pertain even in the event that methods for addressing technical barriers have been put forward

Learnings

The main learnings we took from the project were as follows:

Standardisation

For the organisations we encountered, the work involved in converting data into a standardised format and/or overhauling systems such that all data is recorded in the standard to begin with is extremely onerous, and presents a serious barrier to scaling digital collaborations. An approach which allows partners to get some value from their data without first requiring them to commit time and resources to this effort is therefore of obvious benefit. DDM, for instance, offers a method whereby partners can address queries to a group distributed datasets without requiring them to share the same schema.

It might seem at first as if federated approaches to data collaboration are fundamentally at odds with standardisation projects, but this is not the case, for two reasons. First, a good data standard will always be based on a solid understanding of the needs and priorities of its users - what data means to them, why particular data points are important, and what it will be used for. Federated collaboration is no different. Indeed, it is impossible to design an effective distributed data tool without having first established this base of common understanding. Standardisation and federation share the same social basis.

Second, it may be that in the long run a federated project is the best way to get to standardisation. By lowering the cost of entry to collaboration, a DF could allow partners to build the shared understandings required to undertake more demanding standardisation projects whilst providing them with usable outputs in the meantime. At the very least, anyone committed to developing and implementing data standards across a broad and diverse community ought to consider whether a federated approach could help them achieve their long term goal.

Barriers to Collaboration

Returning to the four core research questions with which we began the project, the first two can now be answered in the affirmative. We were able to engage external partners to scope a project based on the DF model, on the basis that our approach would allow them to achieve something which otherwise would not have possible given their constraints. The premise of our fourth question - that the mitigation of fundamental barriers around privacy, security, standardisation etc would surface other impediments to collaboration - was also proven correct. Ultimately, we were unable to find practical ways to address these "secondary" barriers within the scope of this project, but by documenting them we hope to provide a useful basis for future work in this area.

The purpose of the federated approach to digital collaboration is to provide partners with tools which will allow them to proactively manage issues such as data privacy, security, standardisation and so on. The initial effect of this, we discovered, was not so much to "remove" these barriers as to disaggregate them. In other words, what had previously appeared as a single monolithic barrier (i.e., "data privacy") rendering certain forms of collaboration impossible was translated into a diffuse complex of issues which varied between the different members of the collaboration.

We can list these barriers under the following headings:

- General anxieties around the sharing of data, stemming from a perceived lack of experience and expertise or previous negative experiences
- A lack of clarity around the practicalities and value of the project being proposed
- Perceived or hypothesised differences in goals, motivation or levels of commitment with other partners on the project
- Reluctance to be left with sole responsibility for outcomes or liabilities relating to the project

However, to discuss these issues in abstract terms is to miss the crucial point - the barriers we encountered were highly specific to individual organisations and shaped by their particular past experiences of working with data. These are not problems which can be "solved" through the application of privacy-preserving technologies. Rather, the effect of these technologies was to bring them to light, and indeed, this ought to be understood as part of the primary function of a tool like FL and DDM. By encouraging partners to articulate barriers to collaboration in a more granular and particular manner, they demonstrate what assurances must be provided, what relationships must be built, if the DF is to prosper.

Having surfaced these issues, the next step would be for the DF to discuss them as a collective, with the ultimate goal of producing a signed agreement which would provide partners with the guarantees they needed to feel comfortable as part of the collaboration. As described above, we were unable to steer the DF through this stage during the timeline of the project. The partners preference for signing separate individual agreements with Etic Lab indicates that a major reason for this was the unfamiliarity of our proposal - one-to-one contracts are quite understandably what they are used to and comfortable with. This in turn traces back to two fundamental problems - first, the difficulty of communicating the rationale behind what we were attempting, and second, of demonstrating the value our approach could provide.

The Future

We believe that our work on this project has successfully demonstrated the premise behind Data Federations, and are excited to further develop the theory and practice of federated collaboration. In particular, our next steps will focus on the following issues:

- We've established that DFs show great promise when it comes to mitigating barriers to collaboration, but what collaborative structures and protocols can best motivate participants to engage with the transformative possibilities they offer?

- One of the most appealing aspects of the DF concept to our interlocutors was its relatively low governance burden. Managing distributed data analytics tool is in several respects simpler than a large, complex dataset. Nevertheless, it is still an demanding task, requiring strategic foresight, imagination and the ability to predict and address the wider practical and ethical implications of one's plans - and all of this in a context where some partners will inevitably be more willing and able to take on these responsibilities than others. A more formal accounting of the governance and decision-making mechanisms available to DFs would therefore be of immense value, and will be a priority in the months ahead.

- Thus far, our work has been focused predominantly in the charitable sector. During this project, we appreciated the opportunity to extend our work to an adjacent field (i.e., social finance), but we are keen to expand still further. We believe that the potential applications of the DF concept are extremely wide, but see particular potential for interesting work in local government, unions, co-ops, legal and financial services in particular.

- One of the major hurdles to the implementation of the DF model is the unfamiliarity of the concept itself and the technologies it is based on. We are approaching a point, however, where federated learning and other technologies are going to become a central part of the way that digital services are organised and delivered. Google's so-called "Privacy Sandbox", which includes a proposal to deploy "federated learning of cohorts" as part of a new, "privacy-preserving" approach to targeted advertising, is at the leading edge of this trend. It is paramount, therefore, that public understanding of these technologies is increased, alongside the development of proper standards and regulatory codes to steer their use and implementation. This will be another focus of our work in the next twelve months.

Appendix #1: Building A Data Federation

What follows is a highly-condensed summary of the process of building a DF. It goes without saying that since each DF is a bespoke instance designed to suit the needs of its collaborators, this process will also be customised on a case-by-case basis.

Stage 1: Scoping

A series of workshops and conversations designed to establish the following:

- A common purpose and set of shared goals which can form the basis of a collaborative digital project
- A shared understanding of the available technical options and how they might apply to the goals of the collaboration
- A space for exploring and addressing any risks or ethical concerns which the project might incur
- A register of available resources (both digital and non-digital), and on what terms these can be made available to the DF

Stage 2: Scaling and Governance Agreement

The purpose of this stage is to establish the agreements that will govern the activities of the DF, including:

- A contractual structure setting out rights and responsibilities for all members, addressing liabilities, defining decision-making structures and establishing protocols for the management of any digital tools created and the insights they produce
- A technical specification describing the digital toolset which will be built to address the goals of the DF and the outputs to be generated
- A plan for approaching and onboarding any external parties who have been identified as desirable partners

Stage 3: Deployment

The toolset is built, implemented and results are gathered.

Stage 4: Evaluation and Iteration

The DF reconvenes to evaluate the project, collate learnings and make plans for the future.

Bibliography

Agrawal, N. et al (2021) "Exploring Design and Governance Challenges in the Development of Privacy-Preserving Computation", arXiv preprint. Available at https://arxiv.org/abs/2101.08048. (Accessed: 10 March 2021).

Bunting, M. & Landsell S. (2019) Designing decision making processes for data trusts: lessons from three pilots, Involve. Available at https://www.involve.org.uk/resources/publications/project-reports/designing-decision-making-processes-data-trusts-lessons-three (Accessed: 10 March 2021).

Dahmm, H. (2020) Laying the Foundation for Effective Partnerships: An Examination of Data Sharing Agreements, SDSD Trends. Available at http://www.sdsntrends.org/research/dsainsightsreport (Accessed: 10 March 2021).

Delacroix, S. & Lawrence, N. (2019) "Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance", International Data Privacy Law, 2019, 9 (4), pp. 236-252.

Google (2021) The Privacy Sandbox. Available at https://www.chromium.org/Home/chromium-privacy/privacy-sandbox (Accessed: 10 March 2021).

GovLab (2020) Designing a Data Collaborative. Available at https://datacollaboratives.org/canvas.html (Accessed: 10 March 2021)

Meadway, James (2020) Creating a Digital Commons, London: IPPR.

McMahan, B. & Ramage, D (2017) Federated Learning: Collaborative Machine Learning without Centralized Training Data. Available at https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (Accessed: 10 March 2021).

Open Data Institute (2019), Data Trusts: Lessons From Three Pilots.

Open Data Institute (2020), Designing Trustworthy Data Institutions.

Open Data Institute (2020), Designing Sustainable Data Institutions.

Wark, M. (2019) Capital Is Dead. New York: Verso.