Stanford PACS
Center on Philanthropy and Civil Society
—
Digital Civil Society Lab

Workshop Summary

Trusted Data Intermediaries

Civil society organizations increasingly use a combination of money, time and digital data for public good. The question facing these organizations is how to govern their digital data with the same integrity and alignment to purpose as they govern their financial resources. We are seeing an era of innovation in corporate form and governance as organizations seek to maximize the opportunities of digitized data as a resource, while also ensuring that data are being stewarded and overseen in responsible, public serving ways.

Examples of enterprises trying to do this include Artstor, LearnSphere, the NPC Data Labs, and the Mastercard Center for Inclusive Growth. These are dramatically different organizations managing different types of digitized data, ranging from digital reproductions of art and licensing agreements to student learning data and credit card transactions. What each of these organizations have in common is their commitment to collect, aggregate, and make available large sets of digitized data for public purpose. They each sit between stakeholders, intermediating the relationships between data contributors and data users. In each case, the need to build and manifest trust with these stakeholders and with a broader public is paramount to their success. With this in mind, we are calling these kinds of organizations "trusted data intermediaries."

In December 2016 the Digital Civil Society Lab at the Stanford Center on Philanthropy and Civil Society convened a workshop on trusted data intermediaries (TDIs) to focus on a few key questions:

- What range of organizational forms do TDIs take, how do they work and who do they serve?
- Are they distinct from commercial data aggregators and intermediaries? How?
- What are the common traits and structural manifestations of these organizations? How might we "reverse engineer" a generalizable enterprise form that could serve these purposes in other areas of society?

We identified a small group of sample organizations, including the four mentioned above, and investigated:

- The nature of the data they managed (what it represents, what forms it takes, where it resides, how it is governed)
- The nature of the relationships the organization intermediates (with whom, for what purpose, with what responsibilities)
- The revenue models that support these efforts

With a group of scholars, practitioners, and leaders of similar enterprises in the nonprofit and commercial space, we identified a series of common factors, explored insights about how these organizations work now and might work in the future, and considered the implications of these emerging forms on civil society writ large.

Background

Sometimes the most important innovations can be subtle. Nonprofit organizations in the U.S. now employ almost 10% of the country's workforce, are generally viewed as more trustworthy than other kinds of organizations, and enjoy immensely valuable tax privileges. But at their core, nonprofits are a manifestation of some small but powerful tweaks to the nature of the corporation.

Nonprofits corporations are institutional innovations created to direct privately generated resources toward public benefit. Nonprofits are distinct from their commercial corporate counterparts in two ways. First, they are bound by a non-distribution clause, which directs all revenue above costs to the public purpose of the organization, thus the familiar term "nonprofit," though the term is misleading. The clause doesn't require these organizations to operate without profit; it requires profits to be directed to an organization's public purpose rather than to individuals.

The second distinct design feature of the nonprofit corporation is that it has no shareholders. By replacing corporate shareholders with a public purpose, non-ownership structure, the nonprofit corporate form reinforces the organization's existence to serve a public purpose.

These innovations in corporate code evolved over the past one hundred and fifty years. They have worked well, creating a robust sector of organizations that collect, direct, and dedicate financial and human resources to a public purpose. And as long as those resources have been *rival* and *excludable* economic goods – such as money and time – the model has worked.

However, since at least the mid-1980s we have had the opportunity to add *nonrival* and *nonexcludable* resources to the mix. Specifically, the generation and use of digital data as a resource for public purpose requires us to revisit the nature of the nonprofit corporate form itself. The first nonprofit organizations to focus on digital data or software code as a resource include the Free Software Foundation in 1986, the Electronic Frontier Foundation in 1990, Mozilla and Internet Archive in the mid-1990s. This group began to grow rapidly in the early 2000s with Creative Commons, Wikimedia Foundation, and others.

These early organizations took an existing form – the nonprofit corporation – and used it to manage a new kind of resource, digital data. Trusted data intermediaries are examples of the next iterative step: aligning the organizational form to the resource it seeks to manage. These organizations operate on the premise that digital data:

- Are valuable in the aggregate
- Raise privacy and security concerns;
- Are replicable, networkable, and storable
- Generate new data with every use
- Can be combined with other data sets in a variety of ways

Common characteristics

The first step in our exploration of trusted data intermediaries was to define them. Working from field observations and information collected from workshop participants, we identified the following common attributes.

Trusted data intermediaries:

- Negotiate relationships between contributors and users. Contributors may themselves be aggregators, creating layers of relationships
- Research-specific TDIs depend on Institutional Review Boards as part of their process, often indirectly as IRBs shape the research data collected by the TDI
- The domains of individual privacy and intellectual property are important, and often overlapping, for TDIs
- TDIs work across a spectrum of time – some add value by making timely data available, others add value by preserving digital resources for the long term

- Many TDIs have evolved as a solution to individual organizations' limited capacity for digital storage, preservation, security, and legal undertakings. In so doing, the act of aggregation and intermediation has created new value and opportunities.

There are also numerous important distinctions, including:

- Some TDIs generate revenue, others do not.
- Some focus on being as open as possible while others emphasize control and limiting access
- The data being held may be about individuals or institutions. Similarly, the contributors and users may be individuals and/or institutions.

Suite of negotiations

The suite of issues that TDIs negotiate is perhaps the key distinguishing feature of this phenomenon, as the suite applies across different types of data and across sectors and transcends organizational form. In some cases, it is the province of an entirely new enterprise, in others it is the distinguishing feature of a set of relationships between existing organizations. Some of the most common issues that TDIs negotiate include:

- Ownership, storage, and usage rights for derivative data generated by the use of primary data
- Acceptable standards and practices for identifying data subjects, ensuring anonymity, and algorithmic re-identification
- Access and permissions for analysis
- Responsibility to inform across the layers of relationships, from data subjects to secondary users
- Security requirements
- Tracking and recording the provenance of data and changes in terms over time
- Alignment and communication of multiple sources or licenses
- Regulatory and legal conditions of different institutional partners o *May be able to develop collective liability and responsivity*

Trusted data intermediaries are characterized by this suite of negotiations more so than by a specific organizational form.

Having identified the suite of negotiable issues, we captured the ways in which this work is done. The key question here was: how do these negotiations manifest themselves? We found that there are many mechanisms in use. These include:

- Permissions policies
- Rights management software
- Licenses (CC, CC0, GPL)
- Contract law
- Technological strategies for security, privacy, self-reporting
- Self-regulation and organizational/community review processes, legal and tech review
- Regulatory statutes
- Law (business records regulations, FERPA, HIPPA, COPPA, etc.)
- Self-certification
- Organizational and operational codes of ethics

This is an important set of practices that might – more than any specific organizational form – manifest the answer to managing digital data as a resource for public benefit. The idea that the role of negotiating use and relationships around digital data for public benefit may reside not in an organizational form (such as we've built with nonprofit corporations) but with a set of practices that can be employed within, between, and across organizations, is intriguing.

First, it makes sense in light of the multiple organizational forms, from across sectors, that currently constitute the social economy. Second, practices and mechanisms that can be fitted to specific public

purposes aligns with some of the characteristics of digital data, namely replicability and remixability. In other words, incorporating the social and legal expectations for use into the practices, and not into a single organizational form, matches the reality of both data and sector.

It also sheds light on the ways in which parts of existing organizations might be tweaked or repurposed to fit the possibilities of using digital data for public purpose over time. One idea of such an innovation, a civic trust, is an emerging model. The basic idea is to take the longstanding organizational form of trusts and modify them to address the challenges of digital data. The most important such organizational modification would come in the nature of the trustees themselves, who would be selected to represent both the public purpose interests over time and the communities of data subjects represented in the data set. This proved to be a very useful model, that builds on existing practice, requires minimal changes to existing law, and could be experimented with in the near term.

By focusing on the attributes of the data to be governed, the various mechanisms for governing that can be implemented across organizations, and the potential for making small governance changes in existing organizational forms (the civic trust) we accomplished two things:

1. Clarified the essential elements that characterize a trustworthy approach to using private digital data for public purpose
2. Identified a working set of practices that serve these purposes.

Examples

We worked from four examples: Artstor, LearnSphere, Mastercard Center for Inclusive Growth, and NPC's Data Labs. We also incorporated insights from GitHub and Guidestar. The following section provides brief descriptions of these organizations to illustrate the common characteristics as well as the organizational variation.

Artstor

Artstor is a nonprofit organization that manages access to digital reproductions of artwork in the public domain as well as the various licenses, and rights regimes associated with such artwork. Its primary contributors and users are educational institutions, but it also interacts with rights management organizations, individual artists, nonprofits, and commercial art institutions. Its primary data set is that of the artwork, its secondary data sets include the numerous licensing arrangements it has generated over the years, and its derivative datasets include the statistics on usage of the primary data set. This third set is increasingly valuable as a resource for cultural and educational institutions, curriculum developers, and others. The primary domain of concern is intellectual property and fair use.

Learn Sphere

LearnSphere is a managed dataset of datasets used to teach new learning scientists and to advance the field of learning science. The datasets include digital data generated from students' interactions with various platforms (videos, online curriculum, quizzes) as well as the analytic methods used by the contributing researchers. Users can access both the data and the analytic tools, depending on the access rights negotiated with contributors. All of the data and methods for research have been collected under university-based, IRB-approved research protocols.

Contributors have fine levels of control over allowing access and setting use rights. The data are housed at Carnegie Mellon University. LearnSphere is equally concerned with individual privacy and security issues and intellectual property law as it pertains to scholarly research.

NPC Data Labs

The Data Lab project is focused on allowing UK NGOs a means by which they can assess their programmatic results against UK government-held administrative data. Given the privacy, security and legal liabilities of the government agencies, no publicly held data ever leaves government custody. Instead, NPC Data Labs

negotiates the rights and permissions by which a NGO can provide its dataset to designated government analysts, who compare the programmatic outcomes for a particular population with the outcomes for the larger, matched population. For example, a program working to lower recidivism rates would be able to compare the results of its work on its participants against demographically similar populations for whom the government has data. The analysis is done with government data, on government servers, by government employees and the NGO is provided with the results of that analysis. The NGO data is used only for purposes of evaluation. Access and usage rights are negotiated within the legal and security framework governing UK public datasets. NPC is primarily responsible for encouraging NGO use of this resource.

Mastercard Center for Inclusive Growth

Mastercard, Inc. established an independent arm (MCIG) to facilitate the use of credit card transaction data for public purposes. MCIG negotiates the access and usage rights for researchers (primarily nonprofit and university-based) and manages the creation of and access to a useful, safe, and purpose built dataset. Mastercard's relationship to data on individual card holders is addressed in the cardholder agreements. Negotiations about permissions, access, and use are negotiated individually for each project. Thus MCIG serves as an intermediary between the company and potential research partners. MCIG is primarily concerned with individual privacy, institutional liability, and use rights of researchers.

IP and Trust

The idea of either an enterprise or a set of practices that can facilitate and preserve digital data for public purposes requires us to consider two domains not usually at the center of work on nonprofits or civil society: intellectual property and trust. Intellectual property is the legal domain that has been used to frame (for better or worse) most economic and legal transactions involving digital data. Intellectual property and its legal practices and boundaries - including conceptual and legal work on the commons, public domain, fair use, and the copyleft or alternative licensing movements - has not been a central element of nonprofit law or civil society norms.

From a governance perspective, the increasingly important role of intellectual property law has several implications. IP is itself a framework for thinking about the balance between individual and collective benefit as it generally is used to structure an incentive for individual creation with a longer term benefit to the public. However, the application of IP law to digital artifacts and in digital contexts has been the source of many of the proposed alternatives (creative commons licensing, copyleft movements, other licensing schemes), largely due to the inherent tension between openness and control when easily replicable and mixable artifacts exist in networked environments that allow easy copying and sharing.

Intellectual property law as an influencing domain on digital data intermediaries requires access to a different set of legal knowledge than nonprofits have typically enjoyed. It requires all organizations to consider both the information they create, their means of using it for mission, and the potential financial value of it, in both its original and derivative states. One of the most challenging question for nonprofits as they grow increasingly savvy about digital data as a resource will be whether or not they choose to value their primary and derivative data sets as public purpose outputs or as monetizable assets.

The second domain, trust, is not determined by a body of law. Rather it manifests through – and is nurtured by – sets of behaviors, norms, and relationships. It thus raises a different set of considerations than intellectual property. One of the first distinctions noted here is that efforts to build and maintain trust can be more damaged by regulatory or governing mandates than they may be helped. Organizations that successfully depend on trust, such as libraries (analog) and platforms such as Github (digital) do so largely through clearly communicated norms, transparent processes, user-centric opportunities for engagement, and credible, visible systems of dispute resolution. Trust has to be treated by organizations as something being built and cared for, not something that is achieved and set aside. It is a verb, not a noun. As such, organizations need to consider with whom and about what they are building trust, as there will not likely be a single, one-size-fits-all approach. More important, a real focus on trust is likely to require organizations to require tradeoffs as it becomes selective about who it serves.

Conclusion

By focusing on the suite of negotiations and the mechanisms for manifesting those relationships, the workshop succeeded in articulating several core practices of trusted data intermediaries. Using the identified common characteristics, it would be possible to scan the landscape for emergent forms, to share this information with enterprises considering playing these intermediating roles, and to consider new social domains which might benefit from trusted data intermediaries.

What we did not do was determine if TDIs offer the optimal – or even the only – approach to meeting the needs they currently address. To the contrary, there were several alternative proposals for other ways to facilitate the use of private digital resources for public benefit. These are incomplete, but all build on the sense that the actions we associated with TDIs are actually a suite of responsibilities, manifested through a variety of organizational practices and policies, not necessarily a standalone organizational form. With this in mind the workshop participants briefly considered some alternatives:

- **An accreditation process for the practices, rather than the creation of new organizations.** This might look something like an ICANN or a standards organization, that reviews and accredits the practices and policies for intermediating data. It could also provide a checklist of "how to's," an auditing process for reviewing applicants, and a set of required practices. One proposed requirement was a "continuity plan" that must be created to determine how the data will be protected and made available should the originating organizations/partners cease operations. This approach would be agnostic about the revenue model of the particular enterprise, as long as the model is clearly articulated to stakeholders.
- **The creation of civic trusts for digital data.** These would use the existing legal form of a trust, incorporate the suite of negotiations addressed above, but be distinguished by the composition and responsibilities of the trustees. These individuals, and the process for replacing them over time, would be structured to attend to the long-term protection and use of the data, by and for people representative of those represented as data subjects.
- **A personal information trust** that serves as a digital representation of an individual and their data. This software-based approach to managing rights and access would be controlled by individuals, who would then negotiate access and use of their data on an ongoing basis.

Our investigation into the nature and role of trusted data intermediaries is just beginning.

At this stage three things are clear:

1. The social sector requires new organizational practices to manage digital data for public purpose (and these are emerging)
2. Intermediating data use may be a suite of practices, not necessarily a new set of organizations
3. There are numerous social, legal, and technological approaches to intermediating the use of private data for public benefit.