A Better Deal for Data

The Case for Model Data Commitments and Agreements

Jim Fruchterman, Katy McKinney-Bock, Steve Francis

April 2, 2024

Introduction

We are building a new movement to create a simplified and socially responsible approach to collecting, storing, analyzing, and using data in direct contrast to the surveillance capitalism approaches dominating data processes in today's society. Our goal is a trustworthy set of human-understandable promises backed up by lightweight terms of service and data sharing agreements. We want this to be as close as possible to what the Open Source and Creative Commons movements have achieved in unlocking software code and copyrighted content through a system of widely understood and used open licenses.

We believe that there are many nonprofit organizations, universities, government agencies, and businesses who are looking for a straightforward way to express their commitment to responsible data use which does not exploit the subjects of data collection or their communities. We believe that the full range of institutions from small community-based organizations up to the largest international NGOs, as well as government agencies and businesses who are not depending on monetizing user data, are all groups who should be considering a different approach to data. Furthermore, these organizations would also like to bind technology suppliers to respect their values by making a similar commitment.

We believe that there should be a simple set of enforceable commitments made by all these groups which are understandable by the general public, in stark contrast with the impenetrable and one-sided terms of service which are the status quo today. We are calling this proposed alternative the "Better Deal for Data," and it is the subject of this white paper. This simplified approach is informed by and builds on top of considerable work which has been done by data experts around the world to advance causes such as data stewardship, improved data governance, and responsible use of data in creating AI technologies.

We also believe that the perfect is the enemy of the good. Our assertion is that a standardized set of reasonable and enforceable commitments to responsible data use, which are compatible across organizations and jurisdictions, will deliver benefits across society which will rapidly outpace the current status quo of bespoke agreements and siloed data. We are inspired by the idea of a social license,[i] where society's stakeholders should agree that the uses of data specified in our proposed model agreements are legitimate, and in the overall interests of individuals, communities, global society, and the planet, in contrast with the current model of private value extraction from data.

We hope to build a coalition of the willing to bring about a much wider use of data for the benefit of all of humanity, backed up with the needed guardrails and enforceable promises to use that data ethically and responsibly.

Data is Critical to Social Impact

If you visit a consumer website, 50 different companies will likely be tracking your every move. These companies have a highly developed ecosystem for sharing private data and profiting from it. Search for a pair of shoes and you will shortly start seeing ads for shoes on all sorts of websites!

While the for-profit sector most frequently uses data to intrude into our lives and encourage us to buy things, this technology, when repurposed to focus on equity, can be very powerful for solving a range of social issues in a human-centered manner. This same kind of power is just beginning to be used in the social sector.

Imagine if all the organizations in a city working on homelessness could securely share the private data they collect. These organizations could include community-based organizations, mental health and substance abuse programs, churches, clinics, public safety and others. It would be easier to connect a specific family with the resources that match their needs. You could even align the community towards a new goal of ensuring that every person who was homeless three months ago is now housed. This new approach, "Built for Zero," is already being pioneered in dozens of communities by Community Solutions, which recently won the $100 million MacArthur Foundation 100&Change competition.

The social good sector desperately needs data for doing research, improving programs, and measuring impact. Yet, notwithstanding occasional bright spots like Built for Zero, the nonprofit sector is easily 15 years behind the for-profit sector in applying data to mission-critical needs. The nonprofit sector is barely able to use data to do basic program improvements, much less combine different data sets for maximum insight. This data is highly sensitive, and needs to be handled with care so that it can be used for the benefit of disadvantaged communities and larger society, and protected from those who would use the data for ill or private profit.

Moving to a Social Data Ecosystem that Centers Data Subjects

We need to work together to create the missing social good tech and data ecosystem as a viable ethical alternative to that used by commercial industry to maximize profits. Such an effort needs to be based on the values of equity and ethical action, and avoid the surveillance capitalism imperative to monetize the data of one's users (see our recent SSIR article, "Decolonize Data").[ii] Done right, it is not hard to imagine data innovations doubling or tripling productivity of the people working on socially beneficial activities, or making major progress towards reducing the incidence of a social problem like homelessness. This makes a powerful public benefit case.

This effort needs to be a part of a new social bargain about data between communities and the nonprofits, government agencies, and socially responsible for-profits serving them. We have to promise to keep communities' data safe, use it to their benefit, and not sell them out or use the data

to harm them. We need organizations handling data, who have never really thought critically about their responsibility, to realize that they must handle data with care. And we need legally binding data agreements to guide implementation of a commitment to these promises.

The Need for Model Data Commitments

More and more communities have lost trust in how their data is being collected and used by for-profit companies. A typical farmer is beginning to realize that data about her farm helps the vendors of inputs (seeds, fertilizer) figure out how to charge her more, and the purchasers of her crops figure out how to pay her less. To say the least, these concerns put a damper on data sharing for socially important data collection in agriculture. We've heard from major donors in agricultural research who are worried that society is losing access to crucial data right at a critical moment. To reach society's climate goals, we need data to effectively channel funding to individuals, communities, and organizations which are implementing climate-smart behaviors.

The nonprofits, government agencies, research institutions, and socially responsible for-profits who want to collect, host, analyze, and use data for social impact need a lightweight and standardized solution for building (or rebuilding) trust so that crucial private data can flow to these critical purposes.

We especially need a framework which addresses private data, the kind of data which contains personally identifiable information about individuals or sensitive information about their activities. Nonprofits and agencies regularly collect data from vulnerable people about that which makes them vulnerable. This data deserves a great deal of care, unlike public or aggregated data which can be openly shared.

In some cases, these sorts of challenges around private data have been solved by data trusts or data cooperatives. These are separate legal entities formed to oversee the use of data, typically with a common theme (medical data, public benefits data, farmer data). These are robust, but heavy-weight, solutions which require sophisticated legal preparation and management. They can make sense in wealthy or large economies when multiple large organizations are working together over multiple years. However, setting up a new organization (and governance structure with an independent board of directors) just to handle data is impractical for the great majority of the organizations (and even many local and national government agencies) working around the world on the front lines of social change. A recommendation to establish a separate organization such as a data trust or cooperative is not a match for the great majority of real-world use cases we and our peer organizations encounter. Nonprofits, cooperatives, unions, indigenous communities, and government agencies already have governance structures for accountability to the public or their constituencies. Today's status quo for data governance by these organizations is often non-existent, or has been privatized by tech vendors. We need a simple way for these organizations both to exert control over their data, and commit to manage that data in a socially responsible way, consistent with their existing mission and governance.

Luckily, we already have effective examples of these kinds of solutions from other fields in software, content, and open data. The software community has a range of open software licenses, which are

standardized and well understood, allowing hundreds of thousands of software projects to not worry about legal issues by simply choosing a license that matches their needs. They don't need an attorney to set this up, or explain how the license works. There's already an ecosystem for that.

There is also an excellent model in the open source community from which we hope to learn. The Open Source Initiative has the "Open Source Definition" which sets forth ten criteria that a license must meet to be considered (such as free distribution and no discrimination against people or uses).[iii]

Apart from this special case of software source code, for content of every other kind (photos, videos, research papers, and a wide range of copyrightable materials), we have the family of Creative Commons licenses which make it easy to share creative works with the world under standardized and understandable terms.

The same is true of open data, which "is data that anyone can access, use or share" (The ODI).[iv] It's possible to put a Creative Commons CC0 (public domain) or CC-BY (attribution required) declaration on a dataset and make it open, declaring to the world that it can be used by anyone. Plenty of data deserves to be open, and many governments have committed to open data policies in the interest of transparency. As a result, many interesting datasets are now available, such as the photos of the earth taken by the Landsat satellites, weather data, and aggregated public health data. We are very supportive of the open data movement, and hope that our proposed Better Deal for Data will contribute to the creation of more open data resources derived from data which needs to remain confidential.

Unfortunately, there are no standard approaches for handling private data: data which should not or does not need to be made public. The majority of raw data collected in health, crisis response, education, social services, and so much more, needs to be protected and kept private, especially data which is personally identifiable (easily associated with a specific person). These requirements are often mandated by national law. With the ability to share this data securely with trustworthy organizations, we could be doing far more to directly benefit these communities.

Our Goals for the Model Data Commitments

We originally nicknamed our ideal solution "Creative Commons for Data." We have spoken with Creative Commons and recognize that this isn't literally what we are planning, but it does convey the sense of our goals. And the Creative Commons team has pointed out that their licenses were not designed to be applied to data, even though many groups are doing so today. We think this illustrates the gap we are trying to fill.

Like Creative Commons, we need an easily understandable plain language statement of appropriate commitments about data handling and use. The plain language statements need to be supported by accessible explanations of what they do (and don't) mean plus model legal terms to implement these promises. We are now calling our approach the "Better Deal for Data" (BD4D).

We want it to be simple to apply the Better Deal for Data commitments to a project that collects data—as easy as it is to add a Creative Commons license to a report or a photograph. Not only would it

commit the project to the privacy and security of private data (such as personal information about the subjects of the data being collected) and limit the uses of datasets to socially beneficial ones, but it would require the project to enforce these commitments on other organizations it shares the data with.

One important goal of the BD4D is to go beyond checklists to making enforceable commitments. To generate widespread trust in more socially beneficial uses of data, we believe there needs to be accountability. An organization which is not securely storing sensitive data about vulnerable people (violating their ethical and probably legal obligations) should not make a legally binding commitment to doing so before they have improved their data security. Just as donors in child protection will not fund organizations which lack a child safeguarding policy, we don't believe that donors should fund organizations without data safeguarding commitments like those in the BD4D.

Another goal is to make it much easier to combine data sets from different organizations that are using the BD4D commitments, much like most open source software can be combined with minimal hassles, or several articles with Creative Commons licenses could be assembled for an educational course pack. By adopting these model commitments, it would also become a requirement for commercial tech vendors wanting the business of a socially responsible organization to honor the commitments.

It is possible to spread awareness of better data practices outside the technical and legal community. For example, many European consumers know that there are privacy regulations (the GDPR) which require companies to tell consumers what data is being collected about them, how long it is retained, and their rights in that data (for example, to request errors be corrected). Our model data commitments would build upon understandable approaches like Creative Commons and GDPR to embrace a critical set of data issues.

Initial Draft of the Model Data Commitments

We make the following commitments to "You," all of the individuals or organizations whose data we touch. We make these commitments to You about Your Data which is collected, analyzed, stored, and/or shared:

1. We are using Your Data to benefit You, Your community, humanity, and the planet; not for private gain or profit.
2. We don't claim ownership of Your Data: it remains subject to Your control.
3. We will delete Your Data, correct it, or transfer it somewhere else if You ask.
4. We will not monetize Your Data by providing it to third parties for compensation.
5. You can decide if You want to make Your Data open, or want to monetize it for Your benefit.
6. We will protect and steward Your Data and comply with applicable privacy laws, but You may have privacy obligations as well.
7. If You allow research with Your Data, we will follow best practices around the anonymization of personal data, and published research results will be made available to You for free.

8. We will be bound by legal agreements implementing these commitments, and anyone we share your data with will be similarly bound.

These are just the very earliest drafting of commitments we have considered. We are now embarking on a process of gathering much more input from interested parties. We already have multiple questions we hope to explore, including quite a number from early collaborators. Here are some examples:

- Are there commitments which are missing?
- A commitment to transparency might belong in this list. How should organizations using the BD4D minimize any surprises about subsequent uses of their data?
- Do any of the commitments get in the way of specific good applications of data?
- How sticky should these commitments be on related parties? The obligations are quite strong about direct use of private data including personally identifiable information ("PII"), for organizations accessing "raw" data, but what about research uses for (or AI models trained on) datasets scrubbed of PII?
- Should organizations using software platforms subject to the BD4D commitments (commitments made by the tech provider or a larger NGO hosting the data) be also required to make the commitments?
- Do we need to explicitly deal with downstream use restrictions such as those in various responsible AI licenses, such as prohibiting use of datasets to create weapons?
- More broadly, how should the open source concept of *copyleft* apply to BD4D, if at all?
- Is there a need for some major "flavors" of BD4D, as there are different Creative Commons or open source licenses? For example, Creative Commons has an NC (Non-Commercial) restriction. Should there be something like this in BD4D?
- How do we balance the rights of data subjects to correct or delete data, often under privacy regulations, with the needs of downstream researchers to depend on dataset integrity?
- Clarifying monetization. Every organization needs funding to operate, whether they are a small business, a nonprofit with field programs, or a university interested in doing research. How do we clearly differentiate these "ordinary" ways to find resources to operate from selling confidential consumer data to data brokers and/or large Internet companies such as Meta? Where can we create the bright line between uses which are encouraged by BD4D, and those which are outside the BD4D commitments?
- Assuming that each of the commitments will have an additional level of text explaining what the commitment does and doesn't mean: what needs to be in the explanatory text?
- Traditional privacy approaches focus on the data of the individual, neglecting the interests of groups of people (cancer survivors, members of an indigenous community, etc.). How can (or should) BD4D reflect these community level interests, if the agreements (including the commitments) are often made with, or for the benefit of, individuals?
- Helping nonprofits use and manage data ethically is a major goal of BD4D, but we know that most nonprofits (and many other types of organizations) lack an understanding of these issues (including basic ones such as data privacy and security). What capacity building efforts are

needed from the BD4D effort, to complement the much larger movement to assist nonprofits with tech, data, and AI?

Part of our design process in the next phase of the project is exploring responsible data use cases. For example, we are asking people and organizations to tell us about situations like the following:

- We want to collect [specific kinds of data] from [whom] by doing [specific approach to data collection].
- We want to use that data to [accomplish specific social good] for [specific group of people].
- We want to make sure that [specific bad things] don't happen [to whom] because of the data collection.

We already have received numerous data use cases based on these requests. Two examples are:

- We want to collect [agricultural fertilizer application rate data] from [private landowners] by [requesting they send us their sprayer application data]. We want to use that data to [estimate nutrient inputs] for [their fields] in order to [reduce greenhouse gas emissions]. We want to make sure that [neighbors or environmental NGOs don't sue the landowner for perceived over-application of nutrients] because of the data collection.[v]
- We want to collect [chat conversations] from [children contacting child helplines] by [making a log of text-based conversations that happen online]. We want to use that data to [understand the patterns of child abuse, in order to design better programs to prevent and respond to child abuse] for/on behalf of [children]. We want to make sure that [children are not victimized further] because of the sensitive data collected in service provision.

As we proceed, we want to hear what a wide range of stakeholders would like to see in the commitments and determine if we can build consensus around a strong minimum set of commitments (groups are always encouraged to do more than the minimum!). We want to pressure test the commitments to see if they are fit for purpose by examining a wide range of use cases and challenges.

Beyond the Model Data Commitments

Our next step after revising the model commitments is to support each one with explanatory text to illustrate more of what it means (and doesn't mean) in practice. Single sentence commitments will need more specificity to make them fully useful. Beyond the explanatory text, we plan to craft standardized legal agreements which implement the model data commitments and the associated explanatory text. Creative Commons and open source licenses have shown the benefit of trustworthy, vetted legal agreements which support the larger social goals represented by these movements. Even large sophisticated governments and companies choose to use standard licenses for clarity, compatibility, and wider understanding of their provisions.

In our current vision, a group planning on collecting data and using it for socially responsible purposes would complete a short online questionnaire about these plain language commitments and explaining what data is being collected by whom for what purposes. Then, a tool would deliver the associated data collection, hosting, and/or sharing agreements with enforceable legal terms that implement the

plain language promises. We recognize that it is difficult to translate legal terms into different languages and adjust for different legal systems, but would note that open source licenses such as the General Public License from the Free Software Foundation do not have approved translations to this day, and yet have had world-changing impact.

Can this actually be done? Can we balance simplicity with the fact that data handling is a more complicated issue than the open licensing of copyrighted material like software code, prose, video, or photos? Can we address the fact that context is so important to assessing data uses? We aren't sure that we can, but we think it's important to try!

We Are Not Alone in this Dream

Over the past year, since we were first asked to work on this challenge by Dr. LaKisha Odom of the Foundation for Food and Agriculture Research, we have learned about so many leaders and organizations who see this need or are already working to meet this need in many different fields.

Our team has worked closely with OpenTEAM, a coalition of open source organizations working in the agriculture field in the United States. OpenTEAM has developed an Agriculturists' Bill of Rights, which helped inspire our broader effort here. OpenTEAM has also supported our research into data governance ecosystems.

We would also like to specifically acknowledge that this work builds on extensive and brilliant efforts by organizations such as:

- Ada Lovelace Institute
- Aapti Institute
- Creative Commons
- data.org
- The Data Tank
- Digital Public Goods Alliance
- Global Indigenous Data Alliance
- GO FAIR Initiative
- GovLab at NYU
- Montreal Data License Initiative
- Mozilla Data Futures Lab
- The Open Data Institute (ODI)
- Open Knowledge Foundation
- Responsible AI Licenses
- Societal Thinking

Our goal with the Better Deal for Data is not to replace or supersede these efforts. We are hoping to create a "larger tent" for these efforts to have wider impacts, much like the Open Source Initiative did not replace any of the open source licenses already in wide use, but did create a cohesive identity for most of that movement. Existing data trusts, collaboratives and cooperatives, as well as field and technology-specific initiatives such as the Responsible AI Licenses, are already likely to be shining examples of the Better Deal for Data.

Next Steps

We are in a research phase to learn much more about the needs of our target users, and the existing work to standardize the handling of data, pulling from recent developments around best practices for

sharing in academic science and recent work in machine learning and AI ethics. We already have support from a major Silicon Valley law firm for this effort as a pro bono project. We will be reaching out to many stakeholders in the social sector to get their concerns and use cases. Over the coming months, we expect to provide a steady stream of information about what we learn, as well as early prototypes of what the model commitments, explanatory text, and legal agreements might look like.

Conclusion

We are looking forward to working together with like-minded leaders to develop this idea. Our early conversations have been highly encouraging in convincing us that this is an important, unmet need, and that a workable solution has potential for wide support. We believe that meeting this need is essential to unlocking the full potential of data to serve society in a wide range of fields in the larger social sector. This is much bigger than our nonprofit organization: we will need to build a coalition of the willing. Please let us know if you'd like to get involved!

Acknowledgements

As well as the many, many thought leaders in the data for good field, we particularly want to acknowledge the financial support of the following funders, who encouraged us on this path:

- The Patrick J. McGovern Foundation
- Okta for Good
- OpenTEAM (based at Wolfe's Neck Center, with funding from the USDA)
- Schmidt Futures
- The Skoll Foundation
- Splunk

---

[i] Stefaan G. Verhulst, Laura Sandor, and Julia Stamm, "The Urgent Need to Reimagine Data Consent," *Stanford Social Innovation Review*, July 26, 2023, https://ssir.org/articles/entry/the_urgent_need_to_reimagine_data_consent.

[ii] Nithya Ramanathan, Jim Fruchterman, Amy Fowler, and Gabriele Carotti-Sha, "Decolonize Data," *Stanford Social Innovation Review*, Spring 2022, https://ssir.org/articles/entry/decolonize_data.

[iii] "The Open Source Definition," Open Source Initiative, July 7, 2006, https://opensource.org/osd/. Accessed October 9, 2023.

[iv] The Open Data Institute, "What Is Open Data?," *The Open Data Institute Blog* (blog), September 2, 2016, https://www.theodi.org/article/what-is-open-data/.

[v] Data use case credit to Matt Stephenson of Iowa State University, p.c.