The Common Data Project White Paper v.2

Spring 2011

INTRODUCTION

The Common Data Project is a 501(c)(3) nonprofit organization. Our mission is to expand public access to the sensitive personal information being collected about all of us in a way that balances the privacy rights of individuals and society's need to measure and understand itself.

In the white paper that follows, we describe why we believe personal data needs to be made widely available to the public. We describe the specific data problems that we think a datatrust would solve and the particular challenges the datatrust would have to address to succeed.

Finally, we describe the major components of the datatrust: its basic features as a website and online service and how it differs from other currently available data services. We explain its innovative privacy technology, the community-driven management system, and the governance structure that we feel will best help engender public trust in our project.

TABLE OF CONTENTS

I. ARGUMENT FOR COMMON DATA

The benefits of data, collected from all of us, should be available to all of us. Through a "datatrust," we can create a body of "common data" that could revolutionize research and policy-making; improve transparency; and open the field for data-driven services and products.

A. What is the public getting out of data?

"The future belongs to the companies and people that turn data into products." [1]

Amazon's product recommendation engine, Facebook's social network, Google's targeted advertising platform and a rapidly growing new crop of tech start-ups[3] are all examples of successful transformations of data about people into products. Some of these products are intended for the users themselves. Most often however, the real paying customers of these data products are advertisers. Online sites are most notorious for using data, but credit card companies, grocery stores, and nearly every other major business today rely on data being collected about you and me, who we are and how we behave.

However, it is not clear what "the public" is getting out of all this data collection.

Although Silicon Valley assures us that privacy is a quaint relic of a quickly disappearing world[4], people still obviously care about privacy. Facebook's new features unleashed a PR nightmare last summer, while the Wall Street Journal has dedicated an entire section of its online paper to "What They Know," which describes how online companies are keeping tabs on all of us.[5] More recently, Stanford researchers have put forth a "Do Not Track"[6] technology for browser that is being incorporated by Microsoft and Mozilla into their browsers and legislation has been introduced to Congress.

Advertisers argue that consumers benefit when Amazon recommends a book we might like, or when we get a coupon for something we were thinking of buying.[7] And we certainly understand that advertising pays for "free" services that are truly valuable, like search engines, email, and news.

[1] http://radar.oreilly.com/2010/06/what-is-data-science.html
[3] http://online.wsj.com/article/SB10001424052748704657704576150191661959856.html
[4] http://www.readwriteweb.com/archives/facebooks_zuckerberg_says_the_age_of_privacy_is_ov.php
[5] http://online.wsj.com/public/page/what-they-know-digital-privacy.html
[6] http://donottrack.us/
[7] http://www.nytimes.com/2010/08/30/technology/30adstalk.html

But compared to what companies are getting out of data, what the public is getting out of data is minimal. Increased access to sensitive data could help us understand exactly how tax cuts affect middle class quality of life through direct analysis of credit card receipts, tax returns and bank account activity. Such access could transform the way we understand the society we live in and the choices we make.

However, personal data is largely not available to the public for such uses, at least not on the scale for which it is available to the private sector.

B. What if we made your data your personal property?

Some have argued that one way to deal with the public's lack of involvement in data collection is to give individuals a property right in their own data. That way, they will have exclusive rights to sell their own data and decide how it's used.

This would change a major rule of intellectual property law, which states that no one can own a fact. (This is true in the U.S.; database rights are different in Europe and other parts of the world.)

There are already businesses that operate as if your data is your property. For example, BlueKai[8] and KindClicks collect personal information for market research and provide individuals with a way of managing and monetizing their data. KindClicks, which now seems to be defunct, allowed individuals who contribute data to then donate the money they make off their data to the charity of their choice. BlueKai collects data through cookies, but provides a link on their site by which users can see what information has been gathered about them. Those who want to opt out can. Those that choose to participate in BlueKai's registry can then choose to donate a portion of their "earnings" to charity.

As these businesses illustrate, propertizing data can provide individuals a certain amount of control over their data. It does not, however, give the public access to the kind of valuable data that corporations have. It does not equalize the relationship between large companies and individuals. As BlueKai shows, individuals can opt out, but they can't actively shape the data collection or contribute it to social science research.

Furthermore, it's incredibly difficult to draw a conceptual boundary around personal data and non-personal data. Where would you draw the line? *Your home address. The addresses Amazon.com has delivered to at your request. Your purchase history. Every product page you've ever looked at. The order in which you looked at them. The length of time you spent on each page. What you bought at the end of each browsing session.* And how would we deal with knowledge of facts by more than one person. If I know I'm 33, and you know I'm 33, should I be able to control your use of that knowledge?

Propertizing data in this way fails to take advantage of the most unique and valuable aspects of data: its reusability and its exponentially greater value when used aggregate.

[8] http://bluekai.com

> Data will always be much more valuable in large-scale aggregates to businesses, organizations, government and society as a whole than the individual records will ever be to individuals.

If we made personal data into personal property, there's no guarantee that the individual would gain more power. Individuals already click *"I agree"* without reading terms of use agreements; they could very easily click, *"I sell"* without thinking about it, and receive, at best, a few cents per transaction.

On the other hand, if individuals do want to hold onto their data, the value of each person's data would be so low that individuals might not be offered enough. We could end up in a system in which individuals neither have sufficient financial incentive to sell their data nor sufficient personal use for their data, thereby making it difficult to "mine" personal data because it is too expensive to collate individual records in sufficient quantities.

In both scenarios, the public gets little access to large amounts of interesting data, whether for research or for developing new businesses. The general public, whether researchers or small-scale entrepreneurs, will never be able to compete with large businesses and government if they are required to buy data as well.

Given that the public has provided the data in the first place, we should create a system by which the public doesn't have to pay for access to "their own" data.

C. We should treat data like water, as a public resource.

The Ancient Romans recognized that water was a precious and valuable resource essential to the development of their empire and their society. To ensure that water was available to the public, they developed systems of aqueducts, *castellae* and cisterns, technology ingeniously designed to prioritize water uses and make sure the most important needs of the public were met first.[9] They also developed the doctrine of a "public trust," which states that certain resources are preserved for public use.[10]

Most modern societies continue to manage water as a public resource. Globally, more than 90% of water and sanitation systems are publicly owned and operated.[11]

When water is not carefully managed, the power of private corporations can leave ordinary people without a fundamental resource for sustaining life. In India, the government's groundwater management policy allowed Coca-Cola to deplete the aquifers of Kala Dera, in the arid state of Rajasthan.[12]

[9] Engels, David W. Roman Corinth: An alternative model for a classical city, p. 77.
[10] http://en.wikipedia.org/wiki/Public_trust_doctrine
[11] http://en.wikipedia.org/wiki/Water_privatization
[12] http://www.pbs.org/newshour/bb/asia/july-dec08/waterwars_11-17.html;

We sense a similar problem on the horizon with corporate control over a new kind of precious resource: **personal data**. Because most of this data has been collected as a by-product of transactions between businesses and their customers, the data lies primarily in corporate hands: used, shared, bought and sold between businesses.

It is true that this data would not exist except for the investment they made in hardware, software, human resources and marketing. However, it is equally true that this data would not exist except for the individuals who provided it, oftentimes unwittingly or because they had no viable alternative.

> We are not advocating the creation of a government-run public data utility. We do believe, however, that the enormous value data to society means that like water, data must be made available to the public.

Otherwise, if access to personal data is limited to corporate interests, our society will suffer in the same way a society with poorly managed water systems suffers from the inequities of profit-driven water distribution.

And unlike water, the distribution of data is not a zero-sum game. Unlike water or land, the kind of resources that are normally placed in a public trust, data is easily and infinitely reusable. Thus, providing public access to data will not prevent private businesses from using data as well. In fact, increased general access to data may help the development of more businesses, creating a win-win situation for private and public sectors.

> Therefore, we propose the creation of a datatrust, a publicly accessible store of personal data.

We believe through the datatrust, we can bring the full benefit of the data to the public. Instead of minimizing the value of personal data by dividing it up into individual property rights, we hope to maximize the value of personal data for society by treating it as a shared public resource. Although we are not a governmental entity, as a nonprofit organization, we will allow the public to have access to data that the market does not now provide.

II. PROBLEM SPACE

We have identified particular problems in sharing and accessing personal data that we believe we can solve through the creation of a datatrust.

---

http://www.indiaresource.org/campaigns/coke/

A. Problems of Access to Data

Social scientists and policymakers are lagging behind corporations in their ability to collect, analyze, and access large amounts of data.
Corporations are innovating in leaps and bounds using the wealth of personal information they've collected from millions of customers. Yet researchers and policymakers are still conducting questionnaire surveys with sample sizes in the dozens, hundreds, or if they're lucky, thousands in their effort to answer basic questions about public health, social welfare, the economy and the efficacy of public services. (The Center for Disease Control's National Health and Nutrition Examination Survey (NHANES), one of the most used data sets in public health, collects information from just under 10,000 participants every year.[13]) Even when sensitive data is released, such as with IRS tax returns, only a narrow slice of data is released in aggregate.

> Researchers and policymakers need access to more data to conduct "natural experiments"[14] to answer major policy questions.

- Do tax cuts stimulate the economy? What kind? To what extent? Are the positive effects outweighed by the loss in tax revenue? If so, when?

- How does avian flu spread? Where are there outbreaks? What will the next epidemic be? What containment methods are most effective?

- What factors influence social mobility? Where can the state be most helpful in providing social services? Which social services are ineffective?

- What teaching methods are most effective in helping under-performing students improve? Can they be replicated across the school system? If so, how?

- How is energy consumed in homes and offices? What practices can we change to conserve energy? Where are the greatest potential gains?

- What are the real costs of living, meaningfully adjusted by locale? Are there better ways to measure inflation? What are real-world saving and spending habits? How do they translate into comparable standards of living across income brackets and locales? As a nation, can we come up with a definition of middle-class we can all agree on?

A few large corporations have all the data, making it hard for small players to enter the field.
Although data and data products are some of the most exciting developments in business,

[13] http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/sampling_0708.htm
[14] http://en.wikipedia.org/wiki/Natural_experiment

the vast majority of personal data is held and controlled by a few large corporations. For a robust economy, small-scale entrepreneurs need direct access to data as well.

B. Problems with Releasing Data

Although everyone wants more government transparency, it's hard for governments to release really interesting data.

In the past few years, transparency has become a buzzword for those interested in good and efficient government. The Obama administration launched data.gov. The Sunlight Foundation was founded, proclaiming that it would work to "make government transparent and accountable."[15] New York City and Washington, D.C. announced competitions calling on developers to create useful apps with open city data.[16]

> Yet despite the hubbub, actually finding interesting government data is still quite difficult.

It's difficult and arduous to gather institutional data in all of its various, antiquated formats, collate it, annotate it, and scrub it clean of anything sensitive in order to get it out the door for public consumption. The various open data sites around the country and around the world are changing and improving every day. Still, a quick perusal of the various public data sites reveals a confusing maze of disorganized data files and a paucity of comprehensible, useful data.

For example, the U.S. Census takes years to scrub, swap, and massage their data into "anonymous aggregates" and samples, only to discover that sometimes the data they released is irreparably wrong.[17]

On their website, the IRS has an entire Statistics of Income section devoted to releasing data, but the data released is in predigested aggregates, sampled and massaged by the IRS rather than provided in raw form. Furthermore, because the data must be so heavily "treated" before being released, for some data sets, the IRS is as much as 20 years behind in releasing data.[18]

In order to gain direct access to raw records, you must become one of a handful of top tax policy researchers, go to the Bureau's facility in D.C., browse data on a terminal in a windowless room with nothing more than a pen and paper.

Medicare data available online has gotten richer recently, and sites like Data.Medicare.Gov that allows for interactive access to data. However, to get access to detailed records that contain actual beneficiary-specific and physician-specific information, you must apply for

[15] http://sunlightfoundation.com/
[16] http://www.nycbigapps.com/ ; http://www.appsfordemocracy.org/
[17] http://blog.myplaceinthecrowd.org/2010/02/03/can-we-trust-census-data/
[18] http://www.irs.gov/taxstats/index.html

access and pay for the data as well.[19] The lack of information on the true cost of medical services is certainly contributing to skyrocketing costs—wider access to such information could be crucial in shaping smarter healthcare policy.[20]

In the last decade, there has been a push for nonprofits to be more transparent about the work they're doing. Yet, a dearth of data on nonprofit performance persists.
The difficulty of organizing and sorting data is magnified for nonprofit organizations, whose resources are often quite limited. Many nonprofits struggle simply to file their annual 990 returns.[21]

Organizations like Guidestar have created resources[22] that seek to provide information about nonprofits so you can better identify charities to support. Guidestar has provided a real service in making the information from 990s available in an easy to access, centralized location. The information required on 990s, however, is somewhat limited—the organization's tax status, funding sources and amounts, executive salaries. There is very little public information that provides real data on who nonprofits are serving and how effective they are doing so.

In 2008, Acumen Fund, a venture philanthropy fund released "Pulse,"[23] web-based software that "sets benchmarks for measuring social impact, tracks grantees' performance and compares their performance to other nonprofits doing similar work." It's a big step in the right direction, however the software is intended for proprietary use and the problem of privacy and making data suitable for public release remains an obstacle.

NIH-funded scientists are now required to share their data and findings with the public.[24] Non-profits, which receive a public tax-benefit, should do the same.

It is even impossible for businesses with the resources and the will to release data to enable "open source" analysis.
The companies that collect and store vast amounts of data don't always want to keep it to themselves. In a couple of key instances, these companies have released data to the public, partly as a gesture of goodwill to researchers but also to see if questions they had could be answered effectively through "open source" analysis. However, as the AOL and Netflix experiences showed, there is no safe way for companies to open source data analysis.

In 2006, AOL released 20 million search keywords for 650,000 users over a 3-month period. AOL themselves did not identify users in the report. However, personally identifiable information was present in many of the queries, and as the queries were attributed by AOL to particular user accounts, identified numerically, an individual could be

[19] http://www.cms.gov/IdentifiableDataFiles/
[20] We have conducted a more detailed analysis of different government data sites, which we plan to make available on our website shortly.
[21] http://www.philanthropyjournal.org/resources/managementleadership/preparing-stricter-990-filing-requirements
[22] http://guidestar.org
[23] http://www.acumenfund.org/investments/investment-performance/pulse.html
[24] http://grants.nih.gov/grants/policy/data_sharing/index.htm

identified and matched to their identities and search history by such information, even through simple investigative techniques.[25]

In 2006, Netflix announced a contest with a $1 million prize asking contestants to come up with a better algorithm for their recommendation engine, based on data they had supposedly anonymized. However, two researchers from the University of Texas were able to identify individual users by matching the data sets with film ratings on the Internet Movie Database. Criticism of the Netflix contest, as well as a lawsuit alleging Netflix had violated U.S. fair trade laws and the Video Privacy Protection Act resulted in Netflix cancelling the second stage of the contest.[26]

C. Problems in Public Perception

As stated above, there are many who are struggling to solve problems that require more data. Yet instead of moving towards creating tools to make more data available, we appear to be at an impasse.

The public is increasingly wary of data collection. The companies that actively collect data use "privacy policies" that are meant more to protect companies from liability than protect individuals' privacy rights. Some data collection practices border on fraud and manipulation.[27] Even those businesses that are above-board rarely let the public know the full-scale of what they're collecting.

Furthermore, the public is unsure how data collection benefits them.

More and more, people feel concerned that corporations are intruding into their lives. The standard explanation offered to the public is that data collection is essential to the maintenance and improvement of products and services to customers.

Unfortunately, most businesses have a long way to go in mastering the fine art of balancing intrusion versus utility. Some of the most blatant examples of online tracking are of extremely dubious benefit to the user: e.g., having a pair of shoes we briefly considered on one website continue to stalk us everywhere else we go on the web.[28]

Not surprisingly, the public has responded with fear.

The reason the Wall Street Journal recently launched an entire section devoted to online data collection, ominously called "What They Know," is because the newspaper knew increasing numbers of its readers are wondering precisely, "What *do* they know?"

Public officials have moved to introduce legislation that would restrict data collection and reuse. Although some proposed laws are thoughtful and useful, many simply reflect a fear of the unknown.

[25] http://en.wikipedia.org/wiki/AOL_search_data_scandal
[26] http://en.wikipedia.org/wiki/Netflix_Prize#Privacy_concerns
[27] http://www.nytimes.com/2009/05/17/magazine/17credit-t.html
[28] http://www.nytimes.com/2010/08/30/technology/30adstalk.html

Thus, the debate between privacy advocates and businesses has been framed as a zero-sum game, privacy ***or*** information.

> There is a real danger that our fear of data collection could lead us to "throw the baby out with the bathwater."

III. CHALLENGES TO SOLVING THESE PROBLEMS

We believe that to solve these problems of data access and release, it is not enough to write more detailed privacy policies. We have to make a fundamental shift in the way we provide access to personal, valuable data. Of course, that's easier said than done. There are two major issues that the datatrust must face squarely if it is to solve the problems outlined above.

Public perception is both a problem and an opportunity.
Public fear, misunderstanding, and antagonism towards data collection and data-mining means that any proposal to increase access to personal information will likely stir up justifiable skepticism and some amount of hyperbolic fear. Like the first consumer banks who fought to convince the public that the benefits of handing over your life savings to a faceless institution outweighed the apparent idiocy of doing so, we expect that we will need to create new industry norms to demonstrate public benefit, provide and enforce clear and simple privacy guarantees, clarify personal financial incentives, and provide a credible story for how the organization will stay true to its purported mission.

> There is no such thing as "anonymous." Rather, there are only degrees of privacy and exposure.

The simplest way to address public anxiety about privacy would be to promise 100% anonymity: *Data accessed through the datatrust will be 100% anonymized, full stop.* Unfortunately, complete anonymity is one of the few things we cannot promise, because no one can.[29]

> The word "anonymize" has no widely understood technical definition.

(Similarly, the phrase "personally identifiable information" is defined differently by different countries and laws.[30])

[29] Technically you can guarantee 100% anonymity if you collect no information at all. However, this white paper is clearly the not the place to addressing such an extreme position.

[30] http://en.wikipedia.org/wiki/Personally_identifiable_information

In daily life, we all use the word "anonymous" as if it's a binary concept. You're either anonymous or you're not. But it is impossible to be a hundred percent "anonymous." Rather, privacy is a matter of degree—how much time, effort, and math will it take to reveal your identity?

Unfortunately, the requisite amount is often surprisingly low. So promises of blanket anonymity have to be taken with a giant grain of salt.

De-identification doesn't work. It doesn't protect privacy and it renders data unusable.

Removing "personally identifying information" such as names, social security numbers, addresses, and phone numbers isn't enough to protect identities. This is because seemingly innocuous bits of information when combined can reveal a great deal of information. Latanya Sweeney showed in 1997 that you could positively identify 87% of the population with just a 5-digit zip code, birth date and gender.[31] Law professor Paul Ohm has also written extensively about this problem,[32] and the Netflix and AOL scandals show that these technical problems with de-identification have even broken through to mainstream media.

To make matters worse, the more data there is out there, the easier it will be to deduce your identity by crossing seemingly innocuous data sets. Yet promises of "anonymity" continue to be thrown around freely in privacy policies and user agreements.

Conversely, information that is generally considered to be "personally identifiable information" and therefore scrubbed from data sets could be useful for research by those who have neither need nor desire to know your identity. For example, a researcher doing a survey of popular baby names by socioeconomic status and ethnicity would have no interest in identifying a particular person.

Because of problems with "de-identified data," much data is released as "anonymized aggregates[33]," which we know now may not be truly anonymized.

Last, but not least, with current methods of anonymization, there is actually an inverse correlation between the extent to which you anonymize data and the utility of that data. Paul Ohm argues,

---

[31] http://privacy.cs.cmu.edu/people/sweeney/
[32] http://paulohm.com/
[33] Aggregates are a way of anonymizing data that take individual data records and organize them into predefined reports. (For example, instead of releasing all of the report cards for every 3rd-grader in the school system, you might release a report with counts for the number of students who received an A in each subject, B in each subject, so on and so forth.) The shortcomings of such a technique for releasing data are obvious. Those wishing to make use of such data are limited by the imaginations of the people who designed the aggregate reports. More complex questions like, *"How many students received As in Social Studies, but received Cs or lower in Language Skills?"* are simply not possible.

> "The only way to anonymize a database perfectly is to strip all of the information from it, and any database which is useful is also imperfectly anonymous."[34]

All in all, anonymizing data, as it is done today is not a viable strategy for the datatrust.

For these reasons, anyone seeking to make sensitive data available to the public must demonstrate:

- They can protect private identities in a measurable, enforceable way while giving the public access to private data.
- They can be trusted to truly serve the public interest and not make decisions regarding data based on who is offering the most money.

IV. OUR SOLUTION

> We propose the creation of a datatrust, an online service that makes it easy to release and query sensitive personal information through innovations in community-driven governance and anonymization technologies.

We are the first to acknowledge that the datatrust is an ambitious and complicated undertaking. We've tried to break it down as simply as we can into the following five sections.

In section A. Datatrust Website Features, we provide a laundry list of features and functionality as well as introduce the individuals and organizations we believe will make use of the datatrust and how they will do so.

In section B: CDP and Its Place in the World, we compare and contrast CDP to existing products and services that might be easily mistaken for a datatrust.

In section C. CDP Privacy Technology, we cover the datatrust's central technological innovation, namely our proposal for using differential privacy to provide an objective, measurable and enforceable privacy guarantee.

[34] http://cyber.law.harvard.edu/events/2009/03/ohm

In sections D. CDP Community and E. CDP Governance we answer the questions, "How will the datatrust function?" and "Why should CDP be trusted to run the datatrust?"

A. Datatrust Website Features

The datatrust will consist of a website and web service platform.

The website will function as an entry point for people with data to share and people in search of data to use. Our hope is that it will also grow into a gathering point for collaboration.

Datatrust Stakeholders
Datatrust users will fall roughly into the following 3 categories. Individual users can belong to more than one category.

- **Data Donors:** The government agencies, non-profits and corporations who will donate / release data through the datatrust.

- **Data Users:** The researchers, government agencies, non-profits, corporations and independent entrepreneurs and application developers who will make use of the data made available through the datatrust.

- **Community:** The community of people who will participate in curating and managing the data in the datatrust.

Donating and releasing data to datatrust
Users with data to share will approach the datatrust either to donate data to the datatrust for public use, or because they need to release data to the public and would like to do so through the datatrust. The former can be considered data donors, the latter, data releasers. They will be able to:

- Upload data to the datatrust in a variety of common data formats.

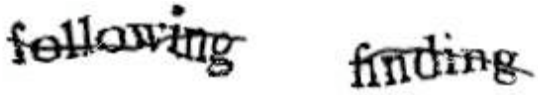
- Create and maintain a profile page for each data set. (More on this below in the "Browsing" section.)

- Create and manage a feed to automatically send new data to the datatrust on a schedule.

Collaborative Data Management.
Data donors and data releasers will need to do the lion's share of preparing their data for public re-use. However, we do imagine that a fair amount of that work can be delegated and distributed to a broader community of volunteers without compromising our privacy guarantee.

Where appropriate, discrete units of data will be distributed through curation tools to individuals to be curated "out-of-context" so that curators don't know where the data is coming from or its significance, with each data point being re-curated by multiple individuals to assure valid categorization and high accuracy. Curating data can be a way for individuals to build reputation within the datatrust community, and can help provide data feed quality ratings of contributors, warn data consumers of un-curated data and assist in prioritizing feeds for future curation.

A good example of this principle at work is Google's clever re-purposing of CAPTCHA[35] technology.



Most of us have come across it when opening a new account. It's a way for websites to prevent spam bots from flooding their site with fraudulent activity by requiring you to decipher and type in a series of letters and numbers that are presumably beyond the capabilities of a computer to recognize; thereby proving that the "user" interacting with the system is indeed a human and not an ill-intentioned spam bot.

In 2009, Google acquired CAPTCHA and since then, they've been using the data from it to help improve the technology they use in Google Books and Google News Archive to convert scanned images of text into text that is data and therefore consumable by their search engines. As a result, Google has been able to harness 5 to 10 seconds of effort from millions of people to contribute to building out their database of print material.[36]

Browsing the Datatrust

Data seekers will be able to browse and search the data.

- **A rich, visual catalog of data profiles** that provide a high-level picture of the scale of the data set (number of records), the demographic range and distribution of the data subjects (gender, age, ethnicity, geography, socio-economic status) and a few key data attributes that highlight the purpose and significance of the data set. Data profile pages will also include a way to examine all the queries that have been submitted to that data set and the answers to those queries.

- **Data projects making use of datatrust data**. This includes where the data is coming from and a composite data profile of that data, as well as the queries that have been submitted and the answers to those queries.

- **Datatrust users and their datatrust contributions.** In addition to their activity, user profiles also include who and what they're following: other contributors, data sets, data projects.

[35] http://en.wikipedia.org/wiki/CAPTCHA

[36] http://techcrunch.com/2009/09/16/google-acquires-recaptcha-to-power-scanning-for-google-books-and-google-news/

Querying the Datatrust
Data researchers will be able to:

- Create and maintain a "data use" project profile page.
- Query a particular data set through a "privacy" filter based on CDP's own implementation of differential privacy.[37] [38]

Community Tools
The community of users will be able to:

- Follow other users, data sets, and data projects.
- View public activity logs for data sets, data projects and users.
- Ask questions and make comments on self-moderated, public forums for data sets, data projects and users.
- Embed datatrust catalog profile pages for data sets, data projects and individual users on external websites.
- Ability to syndicate profile pages and activity logs through various feed mechanisms.

The Datatrust Platform
CDP is not seeking to create *the* datatrust, but the *first* datatrust. We hope to be pioneers in creating a new kind of institution that allows sensitive data to be re-used, and we expect competitors, both for-profit and non-profit, will seek to meet the market's broad needs.

We fully intend for other organizations wishing to set up their own datatrust to be able to license the datatrust technology along with its governance and legal framework.

We also intend to provide a set of APIs to allow others to customize and extend the data curation, querying and data profile functionality and perhaps most importantly, our implementation of differential privacy.

Storage and Security
The data stored in the datatrust will make it a target for attacks, and in order for people to

[37] http://en.wikipedia.org/wiki/Differential_privacy
[38] We will explain the mechanics of differential privacy in more detail in section C: CDP Privacy Technology. For now, it's sufficient to know that CDP is proposing a drastically different approach to making privacy guarantees. The difference can be summed up as a shift away from subjective processes for anonymization towards what we believe to be a more honest model of making transparent the privacy risk incurred by releasing sensitive information, quantifying that risk and making mathematical guarantees around the maximum amount of risk the datatrust will tolerate.

entrust their sensitive information to it, it will need to be properly secured. Though we are familiar with standard industry practices, we are not security experts, and know that this project will require a careful security plan intended to prevent unintended disclosures. While we are looking for an appropriate expert to work with, here are our initial thoughts on some initial ingredients of the plan.

- Data in the datatrust will be stored on servers run and controlled directly by CDP.
- Servers will be housed in a datacenter in caged server racks.
- API[39] access to data filtered through differential privacy will be made available to application developers. Unlike APIs to public data streams like Twitter, datatrust application developers must apply for access with proposals, register with real identities, and will be validated through a human-reviewed, physical mail certification process. API calls must be authenticated and will be logged for review and audit.
- Direct back-end access to raw data will be limited to CDP operations staff. Staff will be background-checked and gain access with two-part, hardware-based authentication for operations pre-approved by the CDP board.

At a more basic level, CDP's security strategy lies in the nature of the service. At heart, CDP is a data catalog under constant surveillance. Data is never allowed to be "checked out" wholesale. Access is limited to logged queries filtered through differential privacy. A user's reputation is developed over time, and every user contact with data is recorded for repudiation and evaluation purposes. There are no opportunities for "Oops, I took it home on a thumb drive and now I can't find it!" scenarios. As a result, the primary external surface area we need to worry about defending is that of intentional, malicious attacks on the service. This may not sound like a big win, but the reality is that most security breaches are the product of accident and human error, not the work of subversive hackers.

Still, cognizant that the datatrust will present an attractive target, and as noted earlier, we are looking for security experts interested in working with us to develop a security plan that meets CDP's unique requirements.

B. CDP in the World: How are we different?

Compared to existing businesses and organizations, what is a datatrust?
There are many services in existence today that look and smell a lot like the datatrust:

- **Data clearinghouses like Bloomberg and Thomson Reuters.** These companies don't just sell stock price data. They're selling any kind of data they can get their hands on. It's why Thomson Reuters suddenly appeared on the market as an unlikely early entry in

[39] http://en.wikipedia.org/wiki/Application_programming_interface

the emerging market for data-driven managed healthcare.[40]

- **A new generation of no-name data brokers.** More obscure companies like Sense Networks who are buying up all of our cell phone data from service providers like AT&T and Verizon. InfoChimps and Spokeo sell data that they aggregate.

- **User-generated data-sharing sites like Swivel** (for any kind of data) and domain-specific sites like PatientsLikeMe and Mint.

- **Government-run public data portals** such as: data.gov, the U.S. Census, NHANES, and a whole range of municipal data sites.

- According to an internal brainstorming document leaked to the media[41], Google is playing around with the idea of becoming a **centralized, open Data Exchange**.

All of these businesses and services collect, collate and release data, either for free or for a fee. But none of them are quite the datatrust we're envisioning.

Unlike for-profit data brokers, our primary stakeholder is the public. We believe the general public is currently under-served by the emerging "data economy."

Unlike the government-run portals, we wish to deal in sensitive, raw, personal information and not just government data.

Unlike the for-profit data brokers who do deal in sensitive data, we will not actually give away any raw, sensitive data. Instead, we only provide a way to query it through a layer of anonymization software.

Unlike any of the data providers described above, we will provide a privacy policy with a measurable definition of what we mean by "anonymized" data.

[40] http://thomsonreuters.com/content/press_room/healthcare/Complete-Audit-Solutions
[41] http://online.wsj.com/article/SB10001424052748703309704575413553851854026.html

C. CDP Privacy Technology: A measurable privacy guarantee.

We've already discussed the public's growing uneasiness about data collection and freewheeling access to sensitive personal information that infringes on their personal privacy. We understood from the beginning that in order to convince the public that what we need is more open access to personal information, not less, we needed to come up with a revolutionary new way to make a privacy guarantee.

We explained in the "Challenges" portion of this paper that currently any and all privacy policies make completely subjective, indefinable, and therefore unenforceable promises around keeping your identity private. Differential privacy is unique in that it provides an objective, quantitative way to measure promises of privacy and anonymity.

So naturally, the next question is, how does differential privacy measure privacy? Not surprisingly, the answer lies in the quantitative way differential privacy anonymizes the data.[42]

When you query data with differential privacy, an unpredictable **amount** of error (a.k.a., noise) is added to the answers to your questions.

You can see this principle in action in this demo we built that allows you to "query" a data set that's been laid out on a map through differential privacy.

It's easy to understand how the introduction of error or noise to answers makes a real difference if you're asking prying questions like, *"How many males born on April 15, 1984, living in zip code: 67201 (Wichita, KS) are HIV+?"* You can easily imagine how the true, raw answer to such a specific question could be 1. You can also easily imagine how even if the true answer were 3 or 13, unambiguous answers at such low quantities would inevitably lead to exposing the identities of the individuals in question.[43]

With differential privacy, you would never get an unambiguous answer such as 3 or 13. You would only get "noisy" answers like $3^{+/-10}$ or $130^{+/-100}$. Still, even with such ambiguous answers, each every answer you get out of a data set will bring you closer to identifying the individuals in that data set. In other words, each and every question incurs a privacy cost. Why? Because even with noisy answers, you can still make statements like, *"With a margin*

[42] http://blog.myplaceinthecrowd.org/2010/05/26/recap-and-proposal-955-the-statistically-insignificant-privacy-guarantee/

[43] Counter-intuitively perhaps, noise and error are necessary to safely answer seemingly harmless questions as well. Even an innocent question like, *"How many kids in Greendale Middle School are over 5'4" tall?"* As re-identification research has shown, data sets can be easily combined and re-combined to reveal information that was not obvious in any one particular data set. Therefore, apparently innocent data about school children and their heights could be re-combined with other data to reveal the identities of sexual abuse victims or some other such deeply sensitive information.

of error of +/-10, there is a 1 in 20 chance that real answer is 3."[44]

> Let's repeat that for emphasis: Even with noisy answers, you can still make statements like, "With a margin of error of +/-10, there is a 1 in 20 chance that real answer is 3.[45]

What is the significance of such statements?
The significance lies in the numbers. We've entered the realm of calculable probabilities of risk. Through programmatic injections of error, differential privacy has given us a way to quantify the amount of information released (a.k.a., the amount of privacy risk incurred) by every question answered by a data set. This in turn allows us to programmatically guarantee that the amount of information a data set releases will never constitute enough information to lead you to draw hard conclusions[46] about individual identities.

> If anonymity is a myth and privacy is a matter of degree, as in degree of personal exposure, differential privacy is a way of calculating that degree.

Differential privacy is exciting, but there are several reasons it is not yet a viable technology.

1. Differential privacy is essentially still a research project.
It's never been deployed in any kind of real user scenarios. It has not been vetted by the broader research community. No one's tried attacking it. No one has tried balancing the utility of the answers from a real dataset against real world privacy requirements. As a result, we don't yet know its practical limitations.

2. Out in the real world, corporations and other groups dealing with privacy issues still perceive privacy to be primarily a Public Relations problem.
These companies and organizations think the privacy problem can be solved with better handling of the media, lobbying legislators and more finely worded policies. In other words, privacy isn't a problem in need of a "technology" solution.

3. The ability to calculate privacy risk is unsettling *and* implies that data resources are finite and can expire.

[44] The statements you can make with differential privacy are actually even more ambiguous than that, mostly because the math behind differential privacy is more nuanced than what we describe here in the whitepaper. However for our purposes of establishing an understanding of the general principles at work, this is an "accurate enough" description of what's going on.

[46] http://en.wikipedia.org/wiki/Statistical_significance/

For the general public, framing data use in terms of "incurring privacy risk" or "increasing personal exposure" is perhaps the quickest way to add fuel to existing anxiety about data collection.

For companies and organizations that collect, hold and use sensitive data, the ability to measure the amount of privacy risk incurred on a query-by-query basis is not necessarily a selling point. Quantifiable privacy risk and privacy policies defined around promises to never incur "too much risk" mean each dataset has a finite lifespan and can only entertain so many queries before it must be retired.

By contrast, today, momentary media flare-ups shine a spotlight on high profile data holders like Facebook and Google, but have little impact on day-to-day data operations. Eventually, the news cycle moves on, "Out of Sight, Out of Mind" takes hold and businesses go about using and passing data around with vague and un-measurable promises to protect individual privacy.

However, we believe that reliance on "Out of Sight, Out of Mind" will eventually backfire in unfortunate ways that will constitute a loss for all of us. For the past year, semi-informed legislators have been pushing for laws that force data collectors to delete data after a certain period of time (e.g., 6 months) not because the data is no longer usable in a safe way, but simply on principle.[47]

Given a choice, we would choose a way to track and calculate when data should be retired based on actual privacy risk over blanket policies that may or may not have any impact on protecting real individuals from actual exposure. However, it is unlikely that the vast, yet still rapidly growing industry of businesses dealing in sensitive data can overcome the inertia of status quo practices to make such a choice.

4. Differential privacy doesn't allow direct access to raw data.
This cannot be sugarcoated: you give up a lot with differential privacy.

Differential privacy would not be useful in many of the contexts in which businesses and organizations currently use data. However, we believe that its limitations are exactly what we need to make the datatrust successful. In other words, for us, it's not a bug, it's a feature.

The biggest limitation of differential privacy is that in order for it to work, there is no direct access to the data. In the ideal, "most private" scenario, as someone trying to ask questions of the data, you are given no information about the data set. You don't know if the set contains 3 people or 3000 people. You don't even know how the data is structured such that each individual record could represent a person a group or an institution.

[47] http://blog.myplaceinthecrowd.org/2010/08/09/in-the-mix-wsjs-what-they-know-data-potential-in-healthcare-and-comparing-the-privacy-bills-in-congress/

This makes differential privacy a non-starter for many who work with data. More often than not, we come to data with only a high-level concept of the questions they would like to ask. It's only after "digging" through it, "getting a feel for it," "playing around with it," that we start to get a sense of what specific questions are interesting to ask. A simple example is that if you have a data set of toddlers, it's pointless to ask questions about dating and sexual activity.

The good news is that if you're working with any significant amount of data, you don't actually ever "look at it" directly to get a feel for it. Public health officials working on NHANES data, for example, don't look at the raw data cells, as it's too overwhelming to provide real information.

So part of what we are investigating is how we can provide researchers with a description of the data (see "A rich visual catalog of data profiles" in the "Our Solution" section V.A.4) so that they can figure out what questions to ask while taking on only a small amount of privacy risk. We think this is a solvable problem.

Still, you could gain a lot with differential privacy.

What excites us about differential privacy is that it might make data available that is unavailable today. The same researchers who are reluctant to have differential privacy stand between them and a dataset might be willing to put up with it if it's the only way for them to gain access to sensitive data. It might be worth it if it allows them to play with data sufficiently to apply for access to raw data; it might be worth it if the data is otherwise uncollectable.

To be clear, we're not advocating for the abolition of working with raw data. There will never be a shortage of situations that warrant such work. However, we also believe there are many situations today where people are working with raw data, not because they need access to individual records, but because there are no viable alternatives; either because anonymization processes take too long, or go too far in narrowing, removing and changing the data.

Additionally, we also see an enormous opportunity in improving the processes and techniques used to release sensitive data to the public today, in terms of the time and human resources required to release the data, the strength of the privacy guarantees the techniques can reasonably uphold, the quality, scope and granularity of the data that is released, and last but not least, the quality of public access to that data.

To illustrate our point, let's take a look at the Census PUMS data.

When data is made available to the public, as the Census does with its PUMs (Public Use Micro-Data Samples)[48], various techniques are used to sanitize the data and make it safe for public release. The techniques usually consist of an ever-shifting, patchwork of manual and

[48] http://www.census.gov/main/www/pums.html

programmatic processes that are both time and labor-intensive, as well as dependent on a lot of subjective "eye-balling it" judgments. They are also largely untested, meaning there have yet to be any serious attacks on Census data to prove the strength of their privacy guarantees. Furthermore, the data samples themselves offer up only a small slice of the total data the Census has at its disposal.

Last but not least, the linchpin of their anonymization process is that it is secret. Meaning, hackers looking to re-identify individuals using PUMs data are prevented from doing so, not because the techniques themselves are immune to reverse-engineering and attack, but because no one on the outside knows how it is done. This approach of "close-sourcing" your security mechanisms is not inherently problematic. The relative effectiveness of open-source versus closed source security is a favorite sparring topic of security experts.[49]

However, in this case, the Census found itself in a bind when it admitted that the accuracy of at least one PUMS was severely compromised by the scrubbing techniques. Why? Because the Bureau could not reverse the scrubbing, as that would then reveal too much about how the dataset was anonymized to begin with.

Differential privacy could allow Census data to be made available without problematic scrubbing. Although it would not be available in the form researchers are most used to using, there would be several clear advantages to the PUMs data.

- Differential privacy is a comprehensive, programmatic system for releasing sensitive information in a safe and responsible way. This means differential privacy is *not* a patchwork, *not* time and labor-intensive, and most importantly, *not* subjective.

- The mechanics of how it works are already available to the public. In other words, the efficacy of the privacy guarantee doesn't depend on secrecy.

- All, as in the sum total, the complete package of data the Census collects, can be made available to the public.

Put yet another way, the choice between the status quo (i.e. PUMS) and the future we propose is not a choice between reliably anonymized samples of data versus noisy, blind access to data. Instead, we are choosing between data that is potentially fraught with irreversible errors, questionably and laboriously anonymized, representing only a slim slice of the total data being collected versus noisy access to the complete data set through a privacy technique that is consistent and transparent.[50]

5. Differential privacy doesn't come with a formula for calculating how much privacy risk is too much privacy risk.

Differential privacy doesn't have a built-in privacy guarantee. It's just a calculator. It doesn't have any opinions or wisdom about when you've incurred too much or too little

[49] http://en.wikipedia.org/wiki/Open_source_software_security
[50] http://blog.myplaceinthecrowd.org/2010/02/05/would-pinq-solve-the-problems-with-the-census-data/

risk. Figuring out how much privacy risk is too much privacy risk is not a simple task, but we've made a fair amount of progress into the problem and believe we see a way out, even if we're not quite ready to offer up a specific number for a privacy budget. Our investigations can be found on our blog.[54]

6. There are many ways to implement differential privacy.
To get into them would require too much math for this white paper. The implementation we've been working with, called PINQ, optimizes for the most accurate within each level of noise. So, even if you decide that a margin of error of +/-100 is good enough for your purposes, PINQ is 3-5 times as likely to give you an answer within +/- 10 of the real answer as it is to give you an answer that is +/-89 of the real answer. As a result, you can quickly burn through a lot of privacy budget even when you don't need to. We're working on ways to re-optimize the curves so that when you say +/-100 margin of error is really good enough, we take you at your word.

At the end of the day, differential privacy enables an "auditable" privacy guarantee.
It's not terribly important that everyone understand the details of how differential privacy works. The important thing to take away is that differential privacy allows for a programmatic way to "quantify" privacy risk, which in turn means that the datatrust's privacy guarantee is built upon a consistent, repeatable methodology that can be audited and validated by people outside of CDP. In other words, when it comes to CDP's privacy guarantee, you won't have to take us at our word.

D. CDP Community: Building a data-sharing community.

We believe that an important part of the datatrust will be the community of data donors, researchers and application developers that emerges to manage and curate the data in the datatrust and work together to find interesting new uses for that data.

We believe that the datatrust should be managed by a community for several reasons:

Turning raw data into useful data takes a lot of work.
As many online communities have already demonstrated (Wikipedia, Freebase, Yelp, Amazon, Facebook photo face recognition), the best way to do that is to enable the people who will get the most out of the data to do that work together.

> We also expect that the community will play a key role in our ability to earn and maintain the public's trust in the datatrust to hold and control large quantities of highly valuable, highly sensitive information.

[54] http://blog.myplaceinthecrowd.org/2010/05/26/recap-and-proposal-955-the-statistically-insignificant-privacy-guarantee/

By relying on the community to run the datatrust, the datatrust distributes power away from the core staff and board and towards the broader user base.

Additionally, we believe that part of earning and maintaining that trust is maintaining a **high level of transparency** about the day-to-day goings-on of the datatrust. The community model will allow us to demonstrate to users how data is contributed, curated, and used in an everyday, functional way. Such a high level of transparency will also allow the general public to understand what the datatrust does, not only what data is available and how people are using it, but also the minutiae of how individual members of the community are engaging and collaborating with one another. We hope the user experience will feel more like Facebook than Wikipedia, where day-to-day edits and discussions are not apparent to the majority of casual readers.

We believe people will join the community, both because they have a need to release or access data, and because they believe in the datatrust's mission. Given the nature of the data we're collecting, we expect that the datatrust will most likely attract organizations and individuals interested in social science issues such as public health, healthcare delivery, education, un(der)employment, urban planning, etc.

We intend to spend significant energy building tools to enable and support the growth of such a community. Although the specifics of the community tools may change, there are certain principles we seek to follow in developing this community.

These principles are derived from research we conducted into a broad range of online communities. We sought to determine practices that will enable the datatrust to thrive and succeed in our mission. In particular, we were interested in how members interact with each other, how they identify with each other and the parent organization, whether they feel responsible for the overall success of the parent organization, whether they feel empowered to govern themselves, what spurs individuals to contribute, and what spurs individuals to care about their reputations in the community.[55]

We fully expect to have to do a lot of listening and adjusting as we go along, but a few things are clear to us upfront.

- Like Facebook and LinkedIn, we will **require real identities** for our members. Participating in the datatrust should be seen as a way to build one's professional reputation.

- High-value data donations, data projects and individual community members need to be **recognized**.

- Members need **easy ways to follow** data, projects and other members of the community.

[55] http://blog.myplaceinthecrowd.org/2010/06/01/ten-things-we-learned-about-communities/

- Data donations, data projects and members should have active profile pages, kept alive with **activity feeds** showcasing recent curation work, queries, questions, requests for peer review, followers etc.

- **Forums** for questions, requests for peer review, requests for new data and general discussion should be **self-moderated**.

- Our **quality evaluation engine** for rating data donations and community contributions should be **driven by user activity**. We feel a straight-forward star-review system won't work well. Instead, evaluations need to be informed by day-to-day usage so that they can be informed by the experience of actual usage and customized to a particular individual's interests and data needs.

- There will need to be **diversity** in the datatrust, By diversity of data, we seek to develop diversity in both the topical content of the data (medical, public health, housing, transportation and urban planning, marketing and retail, financial, etc.) and the demographics of the individuals represented in the data (gender, race and ethnicity, age, geography, socio-economic status, etc.). By diversity of use, we hope to encourage application development, research, *and* policy-making. By diversity of users, we want to attract a wide range of organizations donating and accessing the data (corporations, individual entrepreneurs, non-profits, government agencies, universities and individual researchers).

  We have some early ideas about how best to bring about such diversity, including:

  - Recruiting a diverse board that can help us bring in a broad range of data donations;
  - Providing simple ways for community members to request data

E. CDP Governance

As stated in "Challenges," we recognize that what we need to accomplish is not only technical. We also need to engender the public's trust and confidence that if data is placed with us, we will protect it and make it available only for the public interest, and not for any personal gain.

Although we believe that trust must be earned, we take lessons from other new institutions in history and we plan to signal our trustworthiness in concrete ways. As an online organization, we cannot build stone buildings with pillars and steel-reinforced vaults, but we can invest time and money into an infrastructure that is actually more reliable than bank walls. Ultimately, we may seek legislation similar to FDIC protection for data in datatrusts, but until then, we will implement the following plan to build a trustworthy organization.

We will:

1. Seek out candidates who live and die by their reputations'.
At the end of the day, even the most solidly structured organization will fall down without good people with aligned interests and intentions to run and oversee it. Particularly with board members, we're seeking individuals who have something very real to lose if their reputations are harmed by a data-privacy scandal. Still, we're aware of that "loss of face" is not in itself a fool-proof contraceptive against "bad acts" by "reputable" types, even those who appear to make their entire living off their reputation.[56]

2. Optimize for trust when choosing IRS incorporation status.
Perhaps given the emphasis on "public service," non-profit status is a no-brainer for the datatrust. However, we haven't been ignorant of the recent growth in for-profit social ventures. As a result, we decided to revisit more rigorously our reasons for choosing 501(c)(3) status.

Through our research, we reaffirmed that 501(c)(3) was the right choice for the datatrust for three reasons.

- 501(c)(3) status removes the profit-motive from decision-making.

- Of all the options available to us, 501(c)(3) tax-exempt organizational status invites the most public scrutiny and calls for the strictest tests for whether an organization is truly engaged in serving the public interest.

- 501(c)(3) status invests the public with a sense of ownership and leverage that only "publicly-funded" or government operations can inspire.

We believe that given the particular challenges of trying to be a public repository of sensitive data, the scrutiny and sense of public ownership that come with 501(c)(3) status are crucial to the success of the datatrust and its ability to inspire trust in the public and stay true to its mission.

A more detailed summary of our research, complete with case studies, will be available shortly on our website.

3. Govern through checks and balances.
The datatrust will govern itself through a small staff, a large community, an invested board of directors and a reliance on the public's sense of ownership.

The staff, by virtue of their position will have the most say over day-to-day operational decisions.

The board will be ultimately responsible for any significant changes in mission or policy.

[56] http://www.nytimes.com/2010/01/01/sports/golf/01tiger.html

The community, by virtue of its size and position as the collective workhorse that keeps the datatrust relevant and useful will have the biggest impact on how the datatrust actually serves the public. It will determine that data is available, well-curated and well-used. As a result, any significant changes to the datatrust's goals and processes will automatically require community buy-in.

4. Implement a self-sustaining business model.
We intend to charge a nominal fee for using the datatrust. Initially, the fee will apply only to individuals and organizations seeking access to data. Over time, we imagine that the datatrust could start charging fees for releasing data through the datatrust as well. In this way, we seek to protect the datatrust from the undue influence of a single large-scale funder.

Revenues from fees will go towards operational expenses and staff salaries.

Any major product development work will require additional foundation funding.

Wherever money is involved, there is a danger that the datatrust could become unduly influenced by a single entity at the expense of the public interest. The entity could be corporate, non-profit, institutional, governmental or simply an extremely wealthy, data-hungry individual. Whatever or whoever it is, the datatrust should be prepared to cap the financial influence of any single funding source to some percentage of the organization's total operating budget. The exact number has yet to be determined, but we anticipate starting out at 25% and lowering the cap over time as the datatrust's funds grow more flush.

Now you might ask, *"What if [giant mega-corp] offers you $10 million? Are you really going to turn that down?"*

We strongly believe that we must be careful to convince the public that the datatrust isn't simply a backdoor operation funneling more personal data into corporate coffers. One possible solution is to put that money in escrow and have it trickle slowly into the datatrust as we find matching donations, all the while keeping to our promise of limiting funding from a single source to a pre-defined percentage.

5. Write it all down. Make it public.
We've begun to produce a document that details what the datatrust should be doing and what it shouldn't be doing to serve the public interest with more access to sensitive data.[57] Part mission statement, part ethical treatise, part laundry list of policies and by-laws, this document is not meant to straightjacket the organization. Rather, it is meant to serve as a starting point for ethical and policy discussions and provide the public with concrete leverage when the community begins to worry that the datatrust's staff and/or board members are abusing their positions or leading the organization astray.

[57] http://commondataproject.org/paper-governance-intro

6. Set high standards for transparency.
We've already discussed above the ways in which we think it's important for the datatrust website to be transparent about the activities of the community using the datatrust. We're also compiling a list of organizational metrics we think are important for the staff to share regularly with both the board and the general public. The reports will include fairly standard items such as staff salaries, significant donations, and revenues from fees. However, we also imagine that the staff should be responsible for regularly reporting datatrust "health metrics" such as, volume of donations and usage activity, diversity of donations and data users, rate of curation; volume of un-curated data, and membership growth rates.

7. Implement a "living will."

In addition to the six approaches described above, we will create a "living will" to specify exactly what will happen to your data if the datatrust ever goes bankrupt or is ever dissolved.

a. Protect the raw data.

We will not sell the data. We will not sell the data. We will not sell the data. Under no circumstances will we ever sell the data, nor will we sell the organization to a corporation seeking to use our assets (*aka* the data) for its own profit-driven purposes. This guarantee will be written into our bylaws. However, non-profits can always change their bylaws, but to do so would require upending immense organizational inertia as the community, data donors, the board and the staff would all need to be convinced before such a drastic change could take place. In effect, the only real guarantee we *can* make is that CDP will never be able to sell the data without scrutiny and broader public buy-in.

Responsibly dispose of the data.

There will be a shutdown fund to destroy the data if the datatrust is dissolved and we can't find a new "trusted" entity with a clear public-serving mission to pass the data onto.

V. CONCLUSION & NEXT STEPS

We are fully aware that there is much work to be done to get to the next stage of development. Still, we'd like to take a moment to recognize how far we've come.

We started with a general belief that the struggle between privacy and personal information was perhaps not the intractable zero-sum situation it seemed to be, as well as a vague desire to deliver more meaningful privacy in order to get more sensitive data.

Since then, the world around us has changed.

- **Mainstream media has become increasingly aware of privacy issues,** as illustrated vividly by the Facebook and Google fiascos of 2010.

- **New models of self-monitoring and data-sharing have emerged and become widely popular,** such as Facebook, Twitter, PatientsLikeMe, Mint, FourSquare.

- **The general public has become more aware** of how much personal information is flowing from individuals to corporations and governments.

- And perhaps in response, **an open data movement has begun to claim more data from the government for the people**, emphasizing transparency and government-run data portals.

Since then, we've also done a lot of research and refinement of our work.

In this new world order, **we've identified a small corner of the data-privacy debate we think we can help resolve**, namely that corporations are having all the fun when it comes to getting value out of mining sensitive personal information.

We've identified what stands in the way of creating more access to sensitive data, which in turn allowed us to address those concerns with specific solutions, both with respect to governance and policy and technology.

We've identified the biggest problems with today's privacy policies, namely the lack of an objectively meaningful privacy guarantee, as well as a way to address it.

We've figured out how the datatrust will function on a day-to-day basis through the efforts of an actively engaged community.

The next phase of work we have ahead of us will center around proving we can deliver on our promise of providing an "objective, quantifiable" anonymization technique. This includes:

1. Expanding on the work we've done probing into the math behind the differential privacy guarantee and how that might translate into a privacy promise that is meaningful to laymen.

    a. We built a demo map application with our own implementation of differential privacy.[58]
    b. We also wrote a series of blog posts about what differential privacy means for the end-user.[59]

[58] http://blog.myplaceinthecrowd.org/2011/04/27/the-cdp-private-map-maker-v0-2/
[59] http://blog.myplaceinthecrowd.org/tag/privacy-guarantee/

2. Refining a set of tools we've built to help us visualize how differential privacy works and explain how we plan to use it.

3. With help from the academic community, we would like to re-optimize the differential privacy curves for the specific kind of privacy guarantee we're looking to provide.

4. Creating a Proof of Concept: We think an early application for differential privacy technology would be to build an Anonymous Map-Maker. We could, for example, build an Anonymous Map-Maker for New York City's Department of Health to release in an automated way "anonymized" maps of sensitive disease data to the city's First Responders.

5. Further defining requirements for the back-end datatrust server platform as well as flesh out our security story.

We still have a lot of work to do, but we're excited about the how far we've come and we hope others will join us in supporting our work.

VI. WHO WE ARE

The Common Data Project is a 501(c)(3) non-profit created by Alex Selkirk in the December of 2007 with the intent of exploring alternative models for privacy, data collection and data re-use that would result in the creation of a public data store of sensitive personal information, accessible to researchers, policymakers, individuals and businesses in a safe and private way.

Over the last 10 years of designing, building and using data collection systems, first at Epinions.com and then Microsoft, Alex saw both how valuable data collection was to the companies he worked for as well as the intrinsic limitations of how it was being done. Privacy policies were pitted against data collection in a tug of war that hamstrung data analysis efforts while failing to provide compelling reasons to users to feel good about online services tracking their behaviors.

At the same, Alex became interested in how the data collection techniques being developed in the tech sector could be applied to collecting and releasing data for public use and saw an opportunity to create a more mutually reinforcing model for data collection.

In December of 2007, Alex incorporated the organization and donated the seed money to start the Common Data Project. Grace, Mimi and Geoff joined the team in early 2008.

Alex Selkirk ***(President)*** is currently the principal of SGM, a consulting business he created in 2007 specializing in designing and building large-scale data collection systems and the privacy strategies that go along with them. Alex graduated with a B.A. in Political Science from Yale University.

Tony Gibbon (Contributor, Shan Gao Ma LLC) is a program manager and developer. Tony graduated from University of Washington with a B.S. in Computer Science.

Geoffrey Desa (Board Member) is an Assistant Professor of Business at San Francisco State University focusing on technology social enterprise. Geoff holds a Ph.D. in Entrepreneurship and Strategic Management from the University of Washington, a B.S. in Electrical Engineering from Georgia Tech and an M.S. in Electrical Engineering from Stanford University.

Grace Meng (Emeritus - Education and Communications) is currently a researcher focusing on human rights abuses in the U.S. As a part of her research, Grace travels to different areas of the country to collect first-person stories from U.S. immigrants. Grace holds a J.D. and a B.A. in English from Yale University.

Becky Pezeley (Contributor, Shan Gao Ma LLC) is a veteran Microsoft Program Manager. Becky holds a B.A. in Computer Science, with a certificate in Women's Studies, from the University of Wisconsin - Madison.

Mimi Yin (Board Member) was the Product Designer on the Chandler Project (Open Source Applications Foundation) and currently a graduate student at NYU Tisch School for the Arts - ITP Program. Mimi graduated from Yale University with a B.A. in Music.

You can read our full bios on our website: http://commondataproject.org/about