Mapping Data Flows

A research project on what governs data in society

John Battelle, Juan Francisco Saldarriaga, Matthew Albasi, Natasha Bhuta, Zoe Martin, Veronica Penney

Columbia SIPA
Fall 2018 – Fall 2019

Introduction of the Problem

Data is widely considered to be the oil of the information economy – a core commodity that powers insights, innovation, and wealth creation. But unlike oil, data is non-rivalrous and easily transported and shared. However, the current infrastructure built around data processing has adopted a rivalrous approach – building walls, moats, and other obstructions to the sharing of data across society. This approach has built great wealth for the owners of major data processing firms, but failed to unleash data's potential energy to build a society that exhibits what Edmond Phelps calls mass flourishing.

This failure is the focus of the Mapping Data Flows project.

Hypothesis

Society fails to build what it cannot first envision. Three decades into the digital information revolution, society continues to outsource its imagination to the leaders of the technology industry, which for the most part has adopted a traditional industrial approach to data management. The essential bargain of this approach trades an individual's data for free services based on that data – for example, on today's internet, search results, social media services, news and information services are largely provided at no monetary cost, depending instead on the extraction of personal data as a firm's "fee." The information exchanged between an individual and the service is, for all intents and purposes, locked inside that service's proprietary systems. This ensures that the service has a virtual monopoly on the data it co-creates with an individual, enabling the service to extract maximum profits from the processing of that data.

But what if that data could be shared between competing and complementary services? The Mapping Data Flows project seeks to raise just this question by illustrating the actual architecture of our current data processing system. In its first phase, reported here, the project lays a "cornerstone" by examining and visualizing the most widely adopted policies driving data flows in American society: commercial Terms of Service. Future iterations of the project may extend this initial work across industries and use cases.

The Work

Technology infrastructures and the policies that guide them are maddeningly complex. Visualization is a powerful tool that can clarify this complexity and allow us greater insights. The Mapping Data Flows project has created an interactive multi-layered map of how data flows within one highly complex technology system. Because of its central role in media and social media, the project has focused on the core terms of service for four dominant consumer technology companies in the United States. Research focused on identifying and studying applicable law, firms' end user licensing agreements (EULAs) and Terms of Service (TOS), as well as primary research with subject matter experts and firm representatives. The initial work product is realized as an architectural "blueprint" of how data flows through four key companies' governance architectures.

Methodology, Scope of Data, Companies Studied

Terms of service for major digital players are complex legal documents routinely ignored by the great majority of citizens. They contain confusing and seemingly contradictory statements and often refer to secondary or even tertiary governing policies. A 2012 study found it would take 76 work days for an average American to actually read and understand all the terms of service they encounter in a given year. Amazon alone publishes 14 Terms of Service comprising tens of thousands of words. How could an individual customer possibly understand such a governance framework?

Our first step was to decide which companies to visualize. Nearly every website, app and service has a privacy policy. Even though many policies are similar, they are not similar enough to make sweeping statements about the industry as a whole. Instead, we decided to narrow the focus to the "Big Four" consumer technology companies, because of their scale, scope and name recognition. We also felt readers would be familiar with these services. The project initially focuses on these US-based consumer technology companies:

- Amazon
- Google
- Apple
- Facebook

The MDF project converted each term and data type in a core subset of these companies' policies into discrete database entries. This master database powers an online visualization allowing anyone to explore key insights into the four companies' key policy documents.

For each company's core policy document(s), the MDF team distilled, categorized, inputted, and tagged each pertinent term, in essence converting a static legal document into data. Our goal was to visualize the privacy policies of tech companies to help users understand where, why and which of their data is collected. These privacy policies are a consumer's only window into understanding how their data is collected and used. Our theory is that visualization of this information may make it more comprehensible to the average consumer, as well as useful for

future researchers. In addition, through visualizing this information we hoped to understand how companies differed from one another in their policies and practices.

We focused on a particular subset of these companies' policies. Each of the Big Four has multiple services, products and subsidiaries—each with their own privacy policy and terms of use. Each company also has different terms for its developers and users. The number of interlinking policies is staggering. We began with the main Terms of Use and Privacy Policy for each organization, which are listed below and at the bottom of the visualization at mappingdataflows.com.

After studying each of these policies, we looked for methods of comparison that would work between companies. Since the products and services provided by each company is different, we needed a way to normalize this information. We accomplished this by creating three categories of classification: Data Sources, Data Types and Collection Purpose.

The categories represent Where, What and Why interrogatives. The order of the columns in the visualization is ordered from left to right. Each column, which represents a different category, contains sub-groups of data types arranged in alphabetical order from top-to-bottom.

List of Source Materials

The data behind the visualization was created based on the following terms of service and privacy policies. These source materials do change from time to time, and our research data is based on versions first accessed in January, 2019, and updated through the past year.

Apple Privacy Policy
About Facebook Ads
Facebook Ad Help Center
Facebook Data Policy
Facebook Facial Recognition
Facebook Login and Account Kit
Facebook Payments Inc. Privacy Policy
Facebook Privacy Basics
Amazon Privacy Notice
Google Privacy & Terms
Google Ads and Data Policy
Google Manage Your Location History
Google Payments Privacy Notice
Google One Terms of Service

Rationale for sources not used

After inspecting the Terms of Use for each company, we decided these documents did not address all of the questions we hoped to answer. We decided to focus our data gathering on privacy policies, since these documents explained the company's data collection policies more

clearly. However, due to resource limitations, we did not include the privacy policies for every service each company offers, such as for devices like Amazon's Alexa, or Google's Home or Nest devices.

Data set creation, tools used

We initially coded the data it into neo4j, a graph database. We thought this type of database would be useful in highlighting the connections between types of data, the devices that generate them, the companies that collect them, and what these companies use them for. However, we soon discovered that the actual dataset we created was not complex enough to warrant a graph database. After running through multiple attempts at converting the terms of service into data points, we found that the data was simple enough to reside in a Google Sheet comprised of three tabs, one for the data sources that generate the data, another for the types of data that are generated and collected, and a final one for the collection purpose. Each one of these sheets references the two others. These sheets were then converted into open, downloadable .csv files and make up the basis of the visualization (copies of the files can be found at the base of the visualization).

That being said, our attempts at using the graph database did inform the way we conceptualized the data and how we ended up visualizing it.

Description of visualization tools and rationale for approach

The main library used in the visualization was p5.js which is described as "a JavaScript library for creative coding, with a focus on making coding accessible and inclusive for artists, designers, educators, beginners, and anyone else." The p5.js software allows for a great deal of flexibility and customization. In addition, our team was already familiar with it, which aided in the coding of the visualization.

Initial insights and challenges: Data collection

The data used in this visualization is derived entirely from the information in the main data privacy policies of Amazon, Apple, Google, and Facebook, which are publicly available. We had no inside information or access to company practices. Data privacy policies are first and foremost legal documents. They are extremely carefully written by lawyers for other lawyers and judges, *not* for average consumers, computers, or even data researchers. They are intentionally vague, broad, and difficult to break down. This made translating the policies into a cohesive, tangible data set predictably challenging.

Focusing on one privacy policy at a time, we began by identifying every individual term used and organizing them into categories based on where they appeared in the policy, context, and examples provided by the company. We then focused on the categories that were of most interest to our research: data type, source, and what it's used for. Each company's privacy policy was written independently to suit their own needs, so there was variation in the terms used for similar data. This led to our next challenge, which was to standardize these terms across all four policies in order to represent and compare them in one database. The hundreds

of standardized terms then went through several rounds of consolidation and relabeling, with considerable effort taken to maintain the integrity of the original data.

After extracting the terms to use in the data set, we returned to each privacy policy to connect the dots - where and how each data type is collected, and the various purposes each company can use them for. (Privacy policies often use the term “may” to denote permission, some more than others. In these cases, we included everything a company is allowed to do under its policy, but we can’t definitively say whether they actually do.) Only a limited number of connections are explicitly stated in the policy, such as when Google says “we analyze data about your visits to our sites to do things like optimize product design.”

In order to make most of the connections, the structure and wording of the policies required a fair amount of interpretation. Each policy is primarily split into sections of “this is data we may collect” and “we may use the data we have for.” We had to manually draw inferences from language throughout each policy, often using multiple paragraphs for one data-purpose connection. Additionally, this process was also complicated by the variation of language used by the different companies. Although some standardization of the terms was necessary to create the database, many of the types and purposes are not described in each policy verbatim. Despite these challenges, all of the connections we included in the data set and visualization are supported by the language in the policies. In future iterations of the visualization, we hope to add the actual text for each connection.

Vagueness, standardization issues, nomenclature

As we described above, one of the inherent issues with these privacy policies is that they are very vague. To complicate our process even more, they’re all vague in different ways. Here are a few examples of the problems this raised and how we addressed them:

- The policies we studied have slightly different definitions of what constitutes personal or personally identifiable data. Personally identifiable information (PII) is federally defined and protected, which means that individual companies shouldn’t be allowed to tweak with the definition, but it appears they do. Descriptions of information collected is generally divided into “personal data” and “non-personal data,” although Amazon seems to refer to all the data they collect as personal information - they never mention non-personal information. Facebook explains that some information could be considered “data with special protections,” including religious and political views, which are not necessarily PII. Apple and Google both describe personal data as information that can be used to identify a single person, although they vary in regard to what that entails (for example, Apple considers IP Addresses personal data while Google does not).

  For this project, we relied primarily on the legal definition of PII for the classification of the types of data collected. Any explicit differences between the companies are

represented in the collection purposes, such as what personal data Google is able to share with a third party that Apple is not.

- It is also important to note that what constitutes “products and services” varies between the companies, because it helps explain how data may be used or collected differently. For Facebook the product is mainly the website and apps, including Messenger, Instagram, and Whatsapp. Amazon refers to everything sold through their website as the product, while the platform and features are the services they provide. Apple’s product is their devices, and services are apps such as news, music, podcasts, and access to third party apps from the app store. Google products include both devices, such as Google Home, and platforms like Chrome.

- “Harvested” is a term we came up with to describe data that is constantly being continuously collected, with or without a consumer’s knowledge. None of the companies used the word “harvested” in their policies, but they all mention that they’re collecting it. Amazon has a section about “automatic information;” Apple describes data being collected by “cookies and other technologies;” Facebook details collecting information directly from devices, and Google refers to “information we collect as you use our services.” We found all of these to be either too vague or misleading, so we chose a term that didn’t disguise how the companies were collecting much of our data.

  A lot of the harvested data comes from web cookies that follow us around the internet, but it also includes information such as IP address, current location, and even battery levels. The data that is “harvested” is barely mentioned in the privacy policies, but just from understanding what these companies do, its widely used. We can’t map further than we have from these policies because there just isn’t enough transparency.

Use of language processing software

To help make the task of combing through the text easier, we attempted to use the software Nvivo. We hoped that it would be able to digest our terms and help us extract insights. The initial attempt to use the software allowed us to query and visualize high level questions, such as what was the most used words per document, what were the linkages between conditional words such as “may” and what would come before and after, and anything else that could help us create more transparency around what we were attempting to show. Unfortunately, we wanted to dig deeper than the software allowed. It did allow us to tag text that was vague and not vague and then compare the percentage of text that fell into that category. It turned out that the language used in Apple’s Privacy Policy language was 62.53% vague, Facebook’s was 70.89% and Amazon at 25.57%. We were also able to develop and map our our initial buckets (below) to highlight and tag text blocks and then categorize data into types, usage purpose, source, and the type of language used.

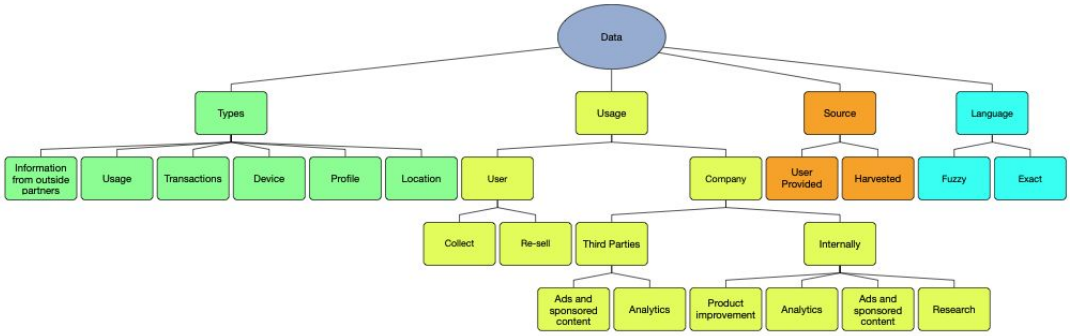

Image: "Source" became the original source of the data (i.e. device, product use, etc), then we used "user provided" and "harvested" to describe how it was collected from the source.

Core insight from early visualizations; legal complexities

We wanted to create a user-friendly visualization that was interactive and could be used to gather insights. Our original conception led us to an organic visualization that presented like a plant-like bud, becoming more complex as one moves through it. However, that metaphor proved too complex for our underlying data set. We had to rework the data set several times based on feedback from the visualization. Our final format allowed us to filter each of the buckets of data sources, types of data, and collection which could then be visualized either in aggregate or filtered by company, case study, collection purpose, or collection method.

Our first version of the visualization based on the data we collected was overwhelming. Even with all of our previous term consolidation, it was still nearly impossible to work through all of the information. We worked diligently to get the visualization to a place where it could be explored more simply, but the fact remains that these companies are collecting an overwhelming amount of data - and they're allowed to use it for almost anything. Their privacy policies are carefully crafted to make this difficult to comprehend, while also ensuring they're legally protected.

These privacy policies were written to hold up in court, and they do. While conducting preliminary research for this project, we found two interesting civil court cases centered around representations made in privacy policies. The first one, also described in one of our insights, involved Apple in 2013. A group of people claimed that "Apple misrepresented its data collection and privacy practices, thereby luring Plaintiffs into spending more money for their iPhones than they would have had they known the true nature of the data being collected by Apple and the third party apps" (p10). They pointed mainly to misrepresentations made in the privacy policy, but Apple's lawyers argued that none of them could prove they actually read the policy. The court concluded that if they never read the policy, they could not have relied on it or been misled by it: "none of these declarations actually states that Plaintiffs read or relied on any particular Apple misrepresentation regarding privacy. Plaintiffs each allude to a vague "understanding"

regarding Apple's privacy policies without providing any evidence whatsoever concerning the basis for this understanding. But a vague "understanding" about Apple's privacy policies is not enough" (p19).

Apple's ability to successfully get the case thrown out on such a technicality highlights much more than the fact that their lawyers are good at their jobs. Apple is relying on all of their customers having only a vague understanding of their privacy practices. By positioning themselves as the most privacy-safe choice when it comes to protecting users' data, they're hoping people won't dive into the details of their policies to find out exactly what they're allowed to access and share. However, even if someone does exactly that, Apple is probably still protected based on a second case we found.

This second case involves the privacy policy of a third party app, but the court's conclusion pertains to all of the companies and policies we researched. In 2018, a judge dismissed a case against UnrollMe because of the way their privacy policy written. He agreed that the customers probably didn't expect the extent of data collection that was being done and that UnrollMe's conduct seemed "unconscionable in the colloquial sense" (p8). Nonetheless, the vague wording in the policy allowed the company to collect, use, and share customer data - *and everyone consented to it.* "Those consumers agreed to the Faustian bargain that undergirds much of the internet: you give me a free service, and I suppress the knowledge that you are probably selling my data to digital touts. We may not like it, but it is not *per se* unlawful" (p9).

Privacy policies are dense, take-it-or-leave-it legal contracts. It doesn't matter that they're difficult to read, vague, or misleading. It doesn't matter if consumers read every word or none at all. That's why a district court judge essentially calling a privacy policy a deal with the devil is an appropriate metaphor - most people don't understand what they're agreeing to, and the price to be paid is higher than anyone expects.

The purpose of privacy policies: To not limit a company's options

It's easy to think of a privacy policy as a document outlining the way that a company protects user information—whether personal data entered to create a profile, content shared on the platform, or financial information from purchases made. That content, however, is only part of the equation. Privacy and data policies also cover the data that companies harvest from users' devices, webpages visited, other apps installed, and mouse or scrolling movements. And these are the most difficult kinds of data to track and to regulate.

Most privacy policies start with a section that broadly describes the type of information the company will collect. User-provided information, such as personal information used to create a profile and content shared on a platform, device information and identifiers, and the websites, third-party integrations, and location information that a company can access while their product is in use fall into this category.

As we've noted earlier, privacy policies are deliberately vague regarding the purposes for which a company can use information. For example Facebook's policy states: "We use the information we have to deliver our Products, including to personalize features and content (including your News Feed, Instagram Feed, Instagram Stories and ads) and make suggestions for you (such as groups or events you may be interested in or topics you may want to follow) on and off our Products."

In this case, "the information [Facebook] has" is all of the information listed in the section detailing the data it collects, and while it may not make logical sense for Facebook to use a user profile photo to personalize content, the privacy policy is broad enough that Facebook *could* use photos for that purpose, if they so choose.

For example, it's fairly simple to restrict GPS location settings on most apps, particularly on iOS devices, but companies can still pinpoint your location through cell phone networks, WiFi networks and IP addresses, or unique device identifiers. If you're signed in on a platform, there's no way to decouple your user information from your devices. Descriptions of how companies use information are often couched in reassuring language, such as, "We use the information we have (including from research partners we collaborate with) to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being. For example, we analyze information we have about migration patterns during crises to aid relief efforts."

Companies also collect information on user behavior, like how long a user spends on a page and where their cursor is hovering. Companies like Facebook use these metrics to make their products more engaging, since the more time a user spends on the site, the more ads they see, and the more money Facebook can make. While privacy policies tend to toward vague language, there is no way to limit Facebook's collection of behavioral information, nor restrict the company from sharing that data with third parties for analytics. This information, which users cannot regulate or protect, also tends to be the information that companies share with third parties and partners.

Case studies

Given how maddeningly vague and comprehensive these company policies can be, the MDF team decided to focus on four specific case studies which deliver core insights about the policies' impact. Each are coded into the visualization and driven by specific data flows from the data set.

The first case study - "Say No Evil, But Keep Your Options Open" - highlights Apple's public stance on privacy versus what its policies actually allow the company to do. The second - "The Illusion of Privacy Settings" - highlights how companies like Facebook give consumers a false sense of control over key information like location data. The third - "Are They Listening?!" - tells a personal story of how data collection practices give rise to creepy triangulation of behavioral

insights. And the fourth - “Absolutely, Definitely Imprecise” - details the vagueness of these policies.

Conclusions, policy implications

Data privacy and the power of technology companies is now a core issue in American and international political dialog. Given this, of the core purposes of the Mapping Data Flows project is to raise awareness around data privacy policies, and to spark conversation about the various regulatory remedies being debated in state, federal and international political bodies.

While the project is still in its early stages, a few things have become quite clear from our work.

- Absent government policy, these privacy policies and terms of service are the core governing documents for how data is controlled in American society.
- These documents are nearly impossible for a typical consumer to read, much less understand.
- The vast majority of consumers do not read these documents.
- The documents are written in an intentionally vague fashion, and to protect the companies they serve, and they are written to stand up in court should they be challenged.
- The documents give an appearance of protection for consumers, but give companies vast control over how data is used, even when “privacy protections” are fully enabled.

Take together, these conclusions seem to support the consensus view of the technology industry as “too powerful” and inherently self dealing. This has led to a wellspring of support of the regulatory remedy of antitrust law. However, our experience with these policies leads us to a different conclusion. While breaking up the big technology companies may limit their powers, it does nothing to change the architecture of control inherent in those companies’ data and privacy policies. As regulators, lawmakers, and policy experts debate the best course forward, we’d suggest they consider novel approaches, including more robust rights related to data sharing (data portability) and the creation of new data sharing market mechanisms. For more on these ideas, see Our Data Governance Is Broken, Let’s Reinvent It (Battelle, 2019).

Next steps, Acknowledgments

We can imagine multiple ongoing initiatives for future iterations of the Mapping Data Flows project. Currently, the data set is static and contained in a simple spreadsheet application. We’d like to map the database to the actual company policies, creating an updated data set that informs a real time visualization.

We would also like to expand our data set to a more comprehensive set of policies, both of the Big Four, as well as across other key industries and companies such as agriculture, health care, and energy. We believe we will find similarities in “non-tech” industries as it relates to the private governance of key data commodities.

We are also eager to engage with a larger ecosystem of data journalists, data visualization experts, and researchers to imagine new approaches to data collection, classification, and visualization. For the next academic year, our work will continue in a more limited fashion, supported both by Columbia SIPA, the Brown Institute for Media Innovation, and the generous support of the Omidyar Network. We would like to acknowledge and thank these organizations for their unwavering support, in particular Dean Merit Janow of SIPA and Professor Mark Hansen of the Brown Institute.

Team Members

John Battelle, Lead. Battelle serves as Adjunct Professor and Senior Research Scholar at Columbia SIPA and project lead for Mapping Data Flows. Founder or co-founder of seven technology and media businesses, author of international best seller *The Search*, investor, board director, and commentator on technology, media, and business.

Juan Francisco Saldarriaga, Visualization Lead. Saldarriaga is the senior data and design researcher at the Brown Institute for Media Innovation at Columbia University where he also teaches data visualization, data journalism, and user interface design.

Mark Hansen, advisor. Dr. Hansen directs the Brown Institute at the Graduate School of Journalism at Columbia University, focusing on statistical analysis, data journalism, and digital media.

Zoe Martin, Research Associate and Masters Candidate, SIPA, Columbia University. Martin has extensive experience in quantitative analysis (Los Alamos Lab) and legal analysis (Vigil Law).

Natasha Bhuta, Research Associate and Masters, SIPA, Columbia University. Bhuta has worked with large corporates and startups on digital product management and strategy.

Matthew Albasi, Research Associate and MS, Data Journalism, Columbia University. Albasi is a documentary filmmaker, local publisher, a specialist in data journalism, and is currently an Investigative Journalism Fellow at Columbia.

Veronica Penney, Research Associate and MS, Data Journalism, Columbia University. Penney is currently an Investigative Journalism Fellow at Columbia.