

How To Read A Privacy Policy

A close reading of 15 online privacy policies as of June 2009. See our [Press Release](#).

The Common Data Project was created to encourage and enable the disclosure of personal data for public re-use through the creation of a technology and legal framework for anonymized data-sharing. Specifically, we think that means creating a new kind of institution called a datatrust, which is exactly what it sounds like: a trusted place to store and share sensitive, personal data.

So why are we spending a lot of time parsing the legalese of some excruciatingly long privacy statements?

We know having an easy to understand, clear-cut privacy policy is critical to the viability of a datatrust. And we felt the first step in figuring out what constitutes an easy to understand, clear-cut privacy policy would be to look at what privacy policies are promising today.

We realize that most users of online services have not and never will read the privacy policies so carefully crafted by teams of lawyers at Google and Microsoft.

And having read all of these documents (many times over), we're not convinced that anyone should read them, other than to confirm what you probably already know: A lot of data is being collected about you, and it's not really clear who gets to use that data, for what purpose, for how long, or whether any or all of it can eventually be connected back to you.

Yet people continue to use Google, Microsoft, Yahoo and more without giving much thought to the privacy implications of giving up their data to these companies.

We at the Common Data Project know that for a datatrust to function properly, we can't rely on people to simply look the other way, nor do we want them to.

Data collection for Google and Microsoft users is incidental. People go to google.com to search, not to give data. As long as they have a good search experience, the data collection is largely out of sight, out of mind.

A datatrust, on the other hand, will be a service explicitly designed around giving and sharing data. We know that to convince the public that the datatrust can indeed be trusted, a clear privacy story is absolutely necessary.

Below you will find a guided tour of privacy policies for 15 online services from established players like Google, Yahoo! and Microsoft to major retailers like Amazon and Ebay, from Web 2.0 starlets like Facebook to aspiring start-ups hoping to compete on superior privacy guarantees. Our goal was to identify when these policies were ambiguous or simply confusing.

Companies Surveyed

The policies analyzed by CDP include those of the companies and organizations listed below. They were picked for being among the most trafficked sites, as well as for providing a range of services online.

Privacy is not exclusively an online issue, even though the companies surveyed here all operate online. Many of the largest data breaches in the last ten years have involved companies and agencies that actually operate exclusively offline, and the question of how to manage, store, and share large amounts of information is an important question for almost every business today. But we chose to

focus on online businesses and organizations because they have been among the most visible in illustrating the dangers, as well as the advantages, of being able to amass great quantities of data.

Search and Internet Portals [Google](#), [Yahoo!](#), [Microsoft](#), [AOL](#)

Major Retailers [Amazon](#), [eBay](#)

Online Communities and Social Networks [Facebook](#),
[Craigslist](#), [Wikipedia](#), [Photobucket](#)

Content Providers [NYT](#), [WebMD](#)

Emerging Search Engines [Ask](#), [Cuil](#), [Ixquick](#)

Here is a quick visual of how their respective privacy policies stack up next to each other, literally.

Questions we asked of each company.

1. What data collection is happening that is not covered by the privacy policy?
2. How do they define “personal information”?
3. What promises are being made about sharing information with third parties?
4. What is their data retention policy and what does it say about their commitment to privacy?
5. What privacy choices do they offer to the user?
6. What input do users have into changes to the policy’s terms?
7. To what extent does they share the data they collect with users and the public?

1. What data collection is happening that is not covered by the privacy policy?

This first question might seem like an odd one. But the fact that there is data collection going on that’s not covered by the “privacy policy” captures so much of what is confusing for users who are used to the bricks-and-mortar world.

When you walk into your neighborhood grocery store, you might not be surprised that the owner is keeping track of what is popular, what is not, and what items people in the neighborhood seem to want. You would be surprised, though, if you found out that some of the people in the store who were asking questions of the customers didn't work for the grocery store.

You would be especially surprised if you asked the grocery store owner about it, and he said, "Oh those people? I take no responsibility for what they do."

Even Walmart, "The Godfather" of business data, probably doesn't let third parties into its stores to do customer surveys that aren't on Walmart's behalf.

But in the online world, that happens all the time. Obviously, when a user clicks on a link and leaves a site, he or she ends up subject to new rules. But even when a user doesn't leave a site, there's data collection by third party advertisers that's happening while you sit there. Companies rarely vouch for what these third party advertisers are doing. Some companies, such as AOL, Microsoft, Yahoo, Facebook, Amazon, and the New York Times Digital, will at least explicitly acknowledge there are third parties that use cookies on their sites with their own policies around data collection. The user is then directed to these third parties' privacy policies. (Note that in the case of New York Times Digital, some of these links are outdated, at least at the time of writing.)

Google, in contrast, doesn't mention third party advertisers on the "privacy policy," alluding to the separate controls for opting out of their tracking on a separate page discussing advertising and privacy.

Companies that don't allow third party advertisers, like Craigslist, of course have no reason to declare this is happening. But most

companies do allow third party advertisers. So an ordinary user with some vague concerns about privacy could decide to finally sit down and read a privacy policy, and then find out that he or she has to read several more policies to really understand who is collecting data, how, and for what. This is an incredible shift from the average user's realm of experience—what grocery store owner would tell a customer to go talk to six different people to understand what was being tracked in that store?

On a related note, many companies have a separate “Terms of Use” which should also be read if a user wants to fully understand his or her rights. For example, when Facebook recently tried to change its terms of use to change its rights to member-generated content, the terms it wanted to change were not in its privacy policy. Yet the privacy of Facebook members was certainly being implicated. So in addition to the various privacy policies that apply to every link, third-party ad, and the site itself, the user must read the terms of use as well.

2. How do they define “personal information”?

Most privacy certification programs, like Truste, require that the privacy policy identify what kinds of personally identifiable information (PII) are being collected. As a result, nearly every privacy policy we looked at included a long list of the types of information being collected.

Many of the companies we surveyed then categorize the information they collect into 1) “personal information” that you provide, such as name and email address, often when you sign up for an account; and 2) cookie and log data, including IP address, browser type, browser language, web request, and page views.

When the first category is called “personal” information, the second category implicitly becomes “not-personal” information. But the

queries we put into search engines are obviously intensely personal.

So are our purchase histories on Amazon, as well as an IP address that can link a certain set of activities to a specific computer.

Yahoo! and Amazon go the extra step of labeling cookie and log data, “automatic information,” giving it a ring of inevitability. Ask Network calls this information “limited information that your browser makes available whenever you visit any website.” Wikipedia similarly states, “When a visitor requests or reads a page, or sends email to a Wikimedia server, no more information is collected than is typically collected by web sites.”

There are companies that do define “personal information” much more broadly. EBay’s definition includes “computer and connection information, statistics on page views, traffic to and from the sites, ad data, IP address and standard web log information” and “information from other companies, such as demographic and navigation information.” AOL states that its AOL Network Information may include “personally identifiable information” that includes “information about your visits to AOL Network Web sites and pages, and your responses to the offerings and advertisements presented on these Web sites and pages” and “information about the searches you perform through the AOL Network and how you use the results of those searches.”

And there are websites that don’t collect information at all: Ixquick and Cuil, the search engines that have been trying to build a brand around privacy. These companies have decided to define “personal” in a rather different way, and in order to protect what is personal, they have chosen not record any IP addresses. Ixquick deletes log data after 48 hours.

We don’t support deleting IP addresses and log data as quickly as possible as a way to protect privacy. We seek solutions for privacy to

preserve the value of data, because we believe that more information is always better than less. But we as a society can't have a thoughtful discussion about what it takes to balance privacy rights against the value of data if companies aren't honest about how "personal" cookie and log data can be.

Some companies do acknowledge that information that they don't consider "personal" could become personally identifying if it were to be combined with other data. Microsoft therefore promises to "store page views, clicks and search terms...separately from your contact information or other data that directly identifies you (such as your name, email address, etc.). Further we have built in technological and process safeguards to prevent the unauthorized correlation of this data." Similarly, WebMD makes this promise: "we do not link non-personal information from Cookies to personally identifiable information without your permission and do not use Cookies to collect or store Personal Health Information about you." WebMD further states that data warehouses it contracts with are required to agree that they "not attempt to make this information personally identifiable, such as by combining it with other databases."

The other companies, however, provide very little explanation of what data combination implies for privacy.

When data is combined, many data sets that initially appear to be anonymous or "non-personally identifiable" can become de-anonymized.

Researchers at the University of Texas in recent years have demonstrated that it is possible to de-anonymize through combination, as when Netflix data is combined with IMDB ratings, or when Twitter is combined with Flickr. So when companies offhandedly note that they are combining information they collect from different sources, they are learning a great deal more about individual

people than the average user would imagine. And as you might imagine, large companies like Microsoft, Google, and Yahoo! have a wealth of databases at their disposal, but none of this is being made explicit in the policies.

3. What promises are being made about sharing information with third parties?

In addition to listing the types of data collected from you, most privacy policies will also list the reasons for doing so. The most common are:

- To provide services, including customer service
- To operate the site/ensure technical functioning of the site
- To customize content and advertising
- To conduct research to improve services and develop new services.

They also list the circumstances in which data is shared with third parties, the most common being:

- To provide information to subsidiaries or partners that perform services for the company
- To respond to subpoenas, court orders, or legal process, or otherwise comply with law
- To enforce terms of service
- To detect or prevent fraud
- To protect the rights, property, or safety of the company, its users, or the public
- Upon merger or acquisition

Nearly every company strives to make these purposes and circumstances sound as standard and normal as possible. “Customize” advertising sounds a lot better than “targeted” advertising, as nobody wants to be a “target.” New York Times Digital even assures its readers that print subscribers’ information will be sold to “reputable companies” that offer marketing info or products through direct mail.

They do also admit that they share information with third parties, but

as inoffensively as possible. Most of the policies we read began their discussion of information-sharing with a declaration that they don't share information with third parties, with the following exceptions. Yahoo states, "Yahoo! does not rent, sell, or share personal information about you with other people or non-affiliated companies except to provide products or services you've requested, when we have your permission, or under the following circumstances." Microsoft similarly promises, "Except as described in this statement, we will not disclose your personal information outside of Microsoft and its controlled subsidiaries and affiliates without your consent." Google's construction is slightly different, but when it states the circumstances in which it shares information, the first circumstance is, "We have your consent. We require opt-in consent for the sharing of any sensitive personal information."

The crucial issue, then, is how "personal information" is defined. And as discussed earlier, the definition of "personal information" varies widely from company to company. When the definition can vary so much, the promise not to share "personal information" isn't an easy one to understand.

For example, Google promises not to share "sensitive personal information," defining it as "information we know to be related to confidential medical information, racial or ethnic origins, political or religious beliefs or sexuality and tied to personal information." Does that mean that a user's search queries for B-list celebrities are fair game to Google? Given the varying definitions of "personal" that are used, the strong declaration that "personal information" will generally not be shared is not, ultimately, a very comforting one.

At the same time, many of these companies admit that they will share "aggregate" or "anonymous" information collected from you. But they

don't explain what they've done to make that information "anonymous." As we know from AOL's experience, a company's promise that information has been made anonymous is no guarantee that it'll stay anonymous.

In this context, Ask Network, in contrast, explicitly lists what it is sharing with third parties, so you don't have to figure out what they consider personal and not personal: (a) your Internet Protocol (IP) address; (b) the address of the last URL you visited prior to clicking through to the Site; (c) your browser and platform type (e.g., a Netscape browser on a Macintosh platform); (d) your browser language; (e) the data in any undeleted cookies that your browser previously accepted from us; and (f) the search queries you submit. For example, when you submit a query, we transmit it (and some of the related information described above) to our paid listing providers in order to obtain relevant advertising to display in response to your query. We may merge information about you into group data, which may then be shared on an aggregated basis with our advertisers.

Ask Network also goes on to promise that that third-parties will not be allowed to "make" the information personal, explicitly acknowledging that the difference between personal and not-personal is not a hard, bright line.

For us at CDP, the issue isn't whether IP addresses are included in the "personal information" category or not. What we really want to see are honest, meaningful promises about user privacy. We would like to see organizations offer choices to users about how specific pieces of data about them are stored and shared, rather than simply make broad promises about "personal information," as defined by that company.

It may turn out that "personal" and "anonymous" are categories that are so difficult to define, we'll have to come up with

new terminology that is more descriptive and informative.

Or companies will end up having to do what Wikipedia does: simply state that it "cannot guarantee that user information will remain private."

4. What is their data retention policy and what does it say about their commitment to privacy?

Data retention has been a controversial issue for many years, with American companies not measuring up to the European Union's more stringent requirements. But for us, it obscures what's really at stake and often confuses consumers.

For many privacy advocates, limiting the amount of time data is stored reduces the risk of it being exposed. The theory, presumably, is that sensitive data is like toxic waste, and the less we have of it lying around, the better off we are. But that theory, as appealing as it is, doesn't address the fact that our new abilities to collect and store data are incredibly valuable, not just to major corporations, but to policymakers, researchers, and even the average citizen. Focusing on this issue of data retention hasn't necessarily led to better privacy protections. In fact, it may be distracting us from developing better solutions.

Google and Yahoo! in the past year announced major changes to their policies about data retention. These promises, however, were not promises to delete data, but to "anonymize" it after 9 months and 6 months, respectively. As discussed previously, neither company defines precisely what the word "anonymize" means. According to the Electronic Frontier Foundation, Yahoo! is [still retaining 24 of 32 digits of users' IP addresses](#). As the Executive Director of Electronic Privacy Information Center (EPIC) [stated](#), "That is not provably anonymous." Yet most mainstream media headlines focused only on the Yahoo!'s claim of shorter data retention. The article in which the

above quote appeared sported the headline: “Yahoo! Limits Retention of Personal Data.”

Interestingly, the debate around data retention has also focused primarily on these three large Internet companies. Even though companies like eBay and Amazon also retain significant amounts of data on their users, there hasn't been any public clamor for Amazon to delete its data as soon as possible. Certainly, the volume and breadth of data Amazon collects pales in comparison to what Google has access to, and some might argue that search queries are more “private” than what books one chooses to buy. But most people still probably wouldn't want their purchase histories on Amazon to be revealed willy-nilly.

A different take on why data retention (which is not addressed at all in its privacy policy) has not become a major issue for Amazon is that Amazon does a better job of showing how its data collection can be useful to its users.

Every item view shows what others have considered buying, what others have ended up buying, what else you might like. In contrast, Google, Yahoo!, and Microsoft have yet to vividly demonstrate why collecting and retaining data makes their services better. Perhaps if they did, they would be less hard-pressed to delete their data as soon as possible.

When I look at a search engine like Ixquick, which is trying to build a reputation for privacy by not storing any information, I'm even less convinced that deleting all the data is a sustainable solution. Ixquick is a metasearch engine, meaning that it's pulling results from other search engines. It's not a solution to replace Google or Yahoo! for everyone. It feels more like a handy tool for someone who is wants to know his search queries aren't being tracked than a model that other

search engines could end up following.

If data deletion by all search engines is the goal, the example to hold up can't be a search engine that relies on other non-deleting search engines!

At the same time, despite all the controversy around data retention, this issue isn't even addressed in the privacy policies of these three large internet companies. Google addressed this issue in a separate FAQs section, while Yahoo! addressed it in a press release and its blog. Microsoft in December 2008 said that they would cut their data retention time from 18 months to six if their major competitors did the same. But this information was not in the privacy policy itself. Among the other companies we looked at, Wikipedia, Ask Network, Craigslist, and WebMd did address the question of data retention in at least some limited way in their policies. No information could be found readily on the sites of eBay, AOL, Amazon, New York Times Digital, Facebook, and Apple.

What exactly do we want to keep private? At the same time, what information do we want to have? What is the best way to balance these interests? These are the questions we should be asking, not "How long is Yahoo! going to keep my data?"

5. What privacy choices do they offer to the user?

Everyone agrees that "choice" is crucial for protecting privacy. But what should the choices be?

Do not call me, email me, or contact me in any way.

Do not let any of your partners/affiliates/anyone else call me, email me, or contact me in any way.

Let me access, edit, and delete my account information.

Let me access, edit, and delete all information you've collected from

- me, including log data.
- Do not track my activities online.
- All of the above.
- None of the above.

Until recently, most tools offered by Internet companies over user information have focused on helping people avoid being contacted, i.e., “marketing preferences.” That’s what we cared about when privacy was all about the telemarketer not calling you at home. Companies have also given users access to their account information, which is in the companies’ interest as well, since they would prefer to have updated information on you.

But few companies acknowledge that other kinds of information they’ve collected from you, like log data, search history, and what you’ve clicked on, might affect your sense of privacy as well. Since they conveniently choose not to call this kind of information “personal,” they have no privacy-based obligation to give you access to this information or allow you to opt out of it.

Still, in the last year or two, there have been some interesting changes in the way some companies view privacy choices.

They’re starting to understand that people not only care about whether the telemarketer calls them during dinner, but also whether that telemarketer already knows what they’re eating for dinner.

Most privacy policies will at least state that the user can choose to turn off cookies, though with the caveat that the action might affect the functionality of the site. AskNetwork developed AskEraser to be a more visible way for users to use Ask.com without being tracked, but as privacy advocates noted, AskEraser requires that a cookie be downloaded, when many people who care about privacy periodically clear their cookies. AskEraser also doesn’t affect data collection by

third parties on its site at all.

More interestingly, Google recently announced some new tools for their targeted advertising program for people concerned about being tracked. These tools include a plug-in for people who don't want to be tracked that will persist even when cookies are cleared and a way for users to know what interests have been associated with them. More information [here](#) and [here](#). Google's new Ad Preferences page also allows people to control what interests are associated with them and not just turn off tracking altogether.

Neither tool is perfect but they're still exciting. The more users are able to see what companies know about them, the better they can understand what kind of information is being collected as they use the service. And Google seems to recognize that people's concerns about privacy can't be assuaged just through an on-off switch, although their controls would be more meaningful if they were more contextual.

Ultimately, however, user choice should result in more fundamental changes to the way data is collected. Right now, Google's targeted advertising program can afford to lose the data they would have tracked from privacy geeks, and still rely on getting as much information as possible from others, most of whom have no idea what is happening. A more significant step forward would be the development of new approaches that will change the way data is collected for everyone.

6. What input do users have into changes to the policy's terms?

Nearly every privacy policy we looked at had some variation on these words: "Please note that this Privacy Policy may change from time to time." If the changes are "material," which is a legal phrase meaning "actually affects your rights," then all data that's collected under the prior terms will remain subject to those terms. Data that's collected after the change, though, will be subject to the new terms, and the

onus is put on the user to check back and see if the terms have changed.

Most companies, like Google, Yahoo!, and Microsoft, promise to make an effort to let you know that material changes have been made, by contacting you or posting the changes prominently somewhere. Some, like New York Times Digital and Facebook, promise that material changes won't go into effect for six months, giving their users some time to find out.

Not surprisingly, nobody states that users will have some say into changes to the privacy policy's terms. Recently, however, [Facebook decided to test out its right to change its terms of use](#),. Facebook wanted to amend the terms of its license to the content provided by Facebook members. Although it wasn't literally a term in the privacy policy, it implicated users' privacy rights as it involved personal content they had uploaded to Facebook. Facebook claimed that its new terms of use did not materially change users' rights but merely clarified what was already happening with data. For example, if user A decides to send a message to user B, and then A deletes her account, the message A sent to B will not be deleted from B's account. The information no longer belongs only to user A. However, Facebook's unilateral attempt to change the terms of use provoked such uproar that the changes were withdrawn. Ultimately, two new documents were created, [Facebook Principles](#) and [Statement of Rights and Responsibilities](#), and users were given the option to discuss and vote on these documents before they went into effect. The new versions were eventually approved by vote of Facebook members.

Facebook may not have set out to become a case study in privacy protection, but this incident is illuminating. Legally, Facebook could change its terms without its members' approval. But practically, it couldn't. There's been some debate over whether the users understood the changes and what they meant, but that's almost irrelevant.

Facebook couldn't simply dictate the terms of its relationship with its users any more, given that its greatest asset is the content created by its users.

It may seem counterintuitive, but it's not surprising that some of the most visible and effective consumer efforts to change how a company uses personal information have come out of Facebook, a service based on voluntary sharing. The more people are given opportunities to participate in how information is shared, the better people can understand what it means for a company to share their information and the more likely they are to feel empowered to shape what happens to their information. Facebook can't offer the service that it does without the content generated by its users. But as it's begun to realize, its users then have to be a part of decisions about the way that content is used.

7. To what extent does they share the data they collect with users and the public?

When we started this survey of privacy policies, our goal was simple: find out what these policies actually say. But our larger goal was to place the promises companies made about users' privacy in a larger context—how do these companies view data? Do they see it as something that wholly belongs to them? Because ultimately, their attitude towards this data very much shapes their attitude towards user privacy.

In the last couple of years, we've seen an unprecedented amount of data collection that's happened largely surreptitiously.

We can't say that we, as users, have gotten nothing in return. The "free" services on the internet have been paid for with our personal

information. But the way the information has been collected has prevented us from negotiating from the vantage point of someone with full information. In other words, we haven't gotten a good deal. The data we've provided is so valuable, we should have struck a harder bargain.

In some ways, consumers are starting to already feel that they've gotten a bad deal. Even though most only feel a vague discomfort at this point, it's unlikely that companies like RealAge will be able to continue what they've been doing. RealAge promoted itself as a simple online quiz to help people be healthier, with endorsements by famous doctors, with only limited disclosure of the fact that their profits were based on selling quiz-takers' information to pharmaceutical companies.

For us at CDP, the fear is that we'll throw the baby out with the bathwater. We don't want to shut down data collection altogether—

We just want companies to stop thinking of our data as their data.

We want to be able to share in the incredible value that this data has, so that we as a society can all benefit from the incredible data collection and analysis capabilities we've developed. Of course, that's only possible with stronger privacy protections than are available now, which is why privacy is such an important issue for us to discuss and understand.

So what would it look like for us to “share” in the value of data? It might sound naïve that companies collecting all this data would ever share data with their users, but it's already happening.

Google, as a company that asserts it's in the business of information rather than advertising, does make some sincere efforts to provide data to the public. [Google Trends](#) may be intended for advertisers, but it also provides the whole world with information on what people

are searching for. [Google Flu Trends](#) is a natural outgrowth of that, and some researchers believe this data can be helpful in determining where flu outbreaks are going to occur faster than reporting by clinics.

Some companies, like eBay and Amazon, have built their data collection into the service they provide to their customers. Some of the information they collect on transactions and ratings can be viewed by all users. Anyone looking to bid on an item on eBay can see how other buyers have rated that seller. A user of Amazon looking to buy a new digital camera can view what other buyers considered.

Although Wikipedia clearly has a different incentive model from the other organizations as a nonprofit, the service it provides also actively incorporates public disclosure of the data collected. The contributions of any one editor can be seen in aggregate and aggregate stats on website activity are also available to the general public. This information is important in the self-policing that is essential for Wikipedia to maintain any credibility.

Although the amount of data these companies are sharing with their users and the public is miniscule compared to the amount of data they've actually collected from us, it raises the possibility that data collection could happen in a completely different way than it does now. Companies could make more obvious that data collection is happening, and rather than frighten users, make the availability of that data another service they provide. The whole process could be one in which users are openly engaged and actively choose to participate, rather than one in which users feel hoodwinked and left out.

Conclusion

By our standards, none of the privacy policies we surveyed quite measure up. Most of them provide incomplete information on what “personal information” means. Many of them fail to make clear that they are actively sharing information with third-parties. Even when they change their policies on something like data retention to placate

privacy advocates, the changes do little to provide real privacy. The legal right companies reserve to change their policies at any time reminds us that right now, the balance of power is clearly in their favor. When they do offer users choices, the choices fail to encompass all the ways online data collection implicates users' privacy.

But we don't believe that we are stuck with the status quo. In fact, there are many positive signs of companies making smart moves, because they're realizing they need buy-in from their users to survive in the long-term. Already, we've seen users trying to determine how their data is shared in all the controversies Facebook has recently endured. Google has created new tools that allow users a wider range of choices for controlling how their data is tracked. And everyday, we see new examples of how data can be shared with users and customers as part of a service, rather than being treated just as a by-product that is solely for the companies' use and enrichment.

We hope that our analysis will help push debate in the right direction. We hope that companies will see there can be real value and return in being more honest with their consumers. At the same time, we hope that as consumers and privacy advocates, we can work with companies towards useful solutions that balance privacy rights against the value of data for all of us.