Privacy Digest

Nutrition Labels for Privacy Policies

David Russell

Table of Contents

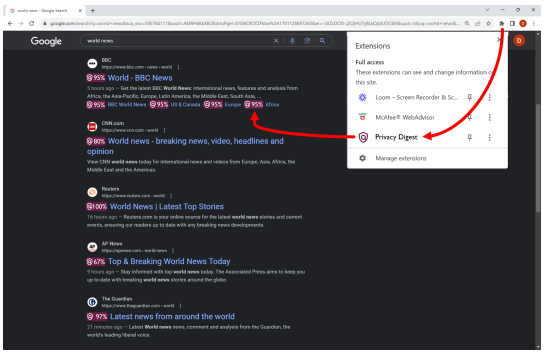
Privacy Digest

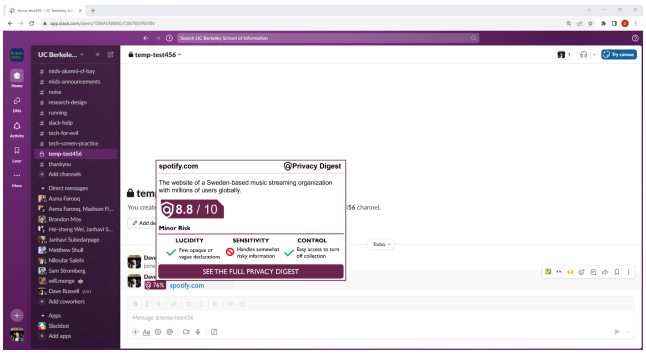
Nutrition Labels for Privacy Policies

David Russell

Overview

Your digital data trail is an increasingly valuable and vulnerable commodity. It is unfeasible to keep track of the information in your web browsing wake and to evaluate the safety of each website you visit. Introducing *Privacy Digest*, a tool to preview the privacy policy of a site before you visit. *Privacy Digest's* team of analysts evaluate policies in the context of the site's purpose and industry, scoring the policy with subject matter expertise and human judgment supported by natural language processing. The strategy is to provide *clear and expedient* evaluation of privacy policies in order to return agency to individuals.





Description

Privacy Digest is a rating system for privacy policies. The goal is to inform individuals of the risks they assume when they consent to a privacy policy and to empower them to opt-out when they have concerns. The system seeks to accomplish this by prioritizing simplicity and ease, condensing the safeguards and risks in a policy into an overall score with three components: Lucidity, Sensitivity, and Control. These scores communicate the risk factor to users while providing click-through functionality to learn the areas of concern in more depth. Scores closer to 10 indicate safeguards for personal information, while scores closer to 0 have riskier factors.

Lucidity	Sensitivity	Control
How clear or opaque is the policy? Are the statements specific? Is it obvious which information is collected, processed, retained, and distributed?	How delicate is the information required by the site? In the context of the use-case, should the site collect the data?	Does the site offer users control over their personal information? Does it default to collecting with opt-out? Or default to not collecting and opt-in?

The three parameters are equally-weighted to derive an overall score. For instance, a website scoring 6, 8, and 7 on Lucidity, Sensitivity, and Control, respectively, has an overall score of 7.

Procedures

To facilitate the team's evaluation of tens-of-thousands of websites, *Privacy Digest* makes use of automation at two stages:

1. Acquisition of privacy policies via web scraping.
2. First-pass assessment of the contents of policies.

Web Scraping

The privacy policies are scraped using Selenium, a Python library for web scraping (Muthukadan 2018). This program relies on contents of privacy policies being available via traditional HTML extraction. As of December 2023, only the main, one-page view of the policy is acquired. In production this functionality would need to be upgraded to accommodate policies that are sectioned into multiple web pages.

Natural Language Processing

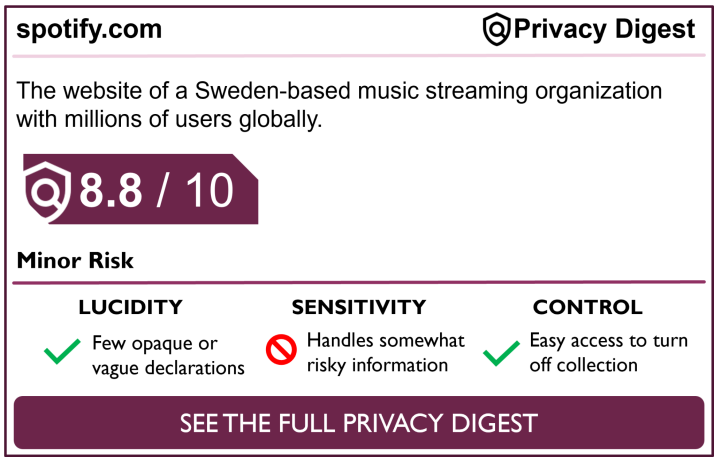
Several procedures are run on the extracted text — checking for risky terms and phrases, detecting precision versus vagueness of language using linguistic inquiry and word count (LIWC 2023), and identifying what data attributes are collected. The automated processes do not assign scores, rather, they consolidate information so that the human reviewer is informed to make judgment calls.

To identify risky terms and phrases, the privacy policy is tokenized and searched for particular strings using regex match. When it finds a match, the program evaluates whether it was negated. For instance, finding a match for "use your data for research purposes" may be preceded by "we do not." The program employs wildcard logic to greedily seek matches, such as searching for the pattern *anonymiz** to find "anonymize," "anonymized," and "anonymization." Wildcards are employed selectively, since they are effective with longer words and more unique spellings but counterproductive in other scenarios such as *prod** matching both "products" and "prodigious." As of December 2023, we search for the following terms and phrases:

- "seek to"
- "strive to"
- "research purposes"
- "purposes necessary to provide products and services"
- "retain your data for as long as necessary"
- "anonymiz*"
- "aggregat*"
- "name*"
- "phone*"
- "email addr*"
- "username*"
- "password*"
- "credit card*"
- "debit card*"
- "Card numb*"
- "bank acc*"
- "physical addr*"
- "street addr*"
- "home addr*"
- "zip*"
- "geoloc*"
- "date of birth"
- "dob"
- "birthda*"
- "birth da*"
- "purchase hist*"
- "transaction hist*"

This key terms and phrases approach expedits the privacy analyst's review of the policy. With strong frontend design it will highlight *where* in the policy the match was found, making discovery of PII statements easy to discover.

User Experience

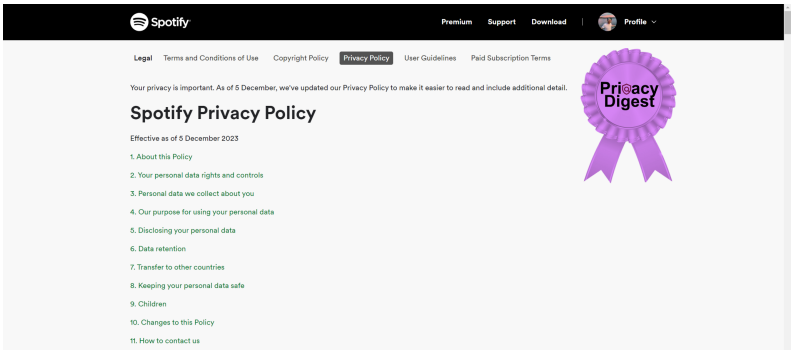


Privacy Digest's success or failure is tied to users' experience of the browser extension and plug-ins. If the score is intuitive and the underlying parameters are cogent, then users will be more likely to continue use. The modal, which is designed to be universal across extensions and plug-ins (e.g., Chrome, Bing, DuckDuckGo, Slack, etc.), prioritizes the overall score, placing it in the upper-left corner where users' eyes are drawn first. It is in larger font and bolded. The individual scores, Lucidity, Sensitivity, and Control are featured next, lower in the modal and in smaller font. The final key element of the modal is the *See The Full Privacy Digest* button, which provides click-through functionality for users to read a thorough report. While this button is large, occupying a large percentage of the modal surface area, it is positioned last, and is therefore deprioritized versus the scores. This strategy is intentional. While ideally all users would seek out holistic understanding of sites before they visit, it is unlikely that most users will spend more than a few seconds pause before proceeding. *Privacy Digest's* approach makes the conscious choice to prioritize judgments based on score over thoroughly educating users.

Adoption Strategies

Privacy Digest operates as an independent entity with scores assigned based on internal employee appraisals of privacy policies. The score-assignors will do so independent of the organization they are evaluating. This strategy has the benefit of being an independent audit, above reproach. But it also has the disadvantage that the policies do not tell an organization's entire data usage story.

An alternative approach to consider is organization opt-in. *Privacy Digest* could be released as an opt-in product, only scoring the privacy policies of organizations that agree to be evaluated. This would translate the operations to a collaborative approach, with privacy analysts coordinating with organization representatives to understand their policies, working with them to make adjustments, and ultimately publishing a score that *is influenced by the organization*. This approach has the benefit of reducing legal action and disputes. But it has the drawback of potentially losing user trust if they consider the scores to be fluff, disguised propaganda issued by organizations through a third-party.



Getting organizations to opt-in before *Privacy Digest* has credibility would be challenging. One way they could be incentivized to participate is with tactics like a *kudos-badge*, drawing attention to their openness to external scrutiny.

Example: Spotify and Apple Music

To clarify the above concepts and design, here is a cursory example comparing the privacy policies of Spotify and Apple Music. This approach can accurately be characterized as qualitative judgments supported by quantitative evidence.

A few initial observations:

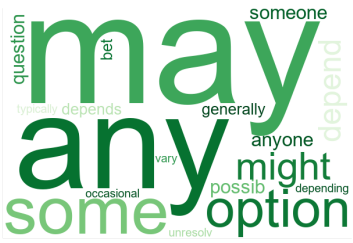
- Spotify's policy contains approximately 5,400 words, versus 3,200 for Apple Music.
- Spotify mentions many more PII attributes specifically than does Apple Music.

Contains 11 flagged keywords:	Contains 4 flagged keywords:
aggregat*, card numb*, credit card, date of birth, email addr*, name, password, phone, street addr*, username, zip	aggregat*, email addr*, name, phone

Lucidity

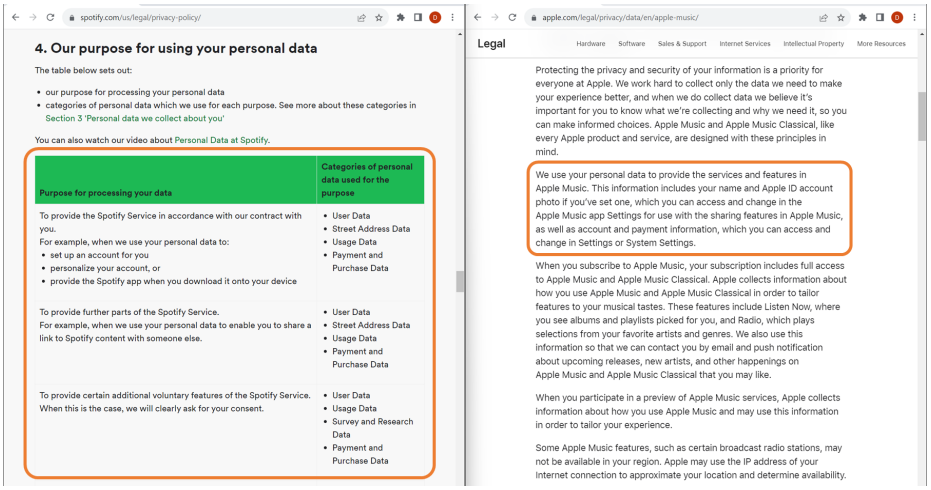
Word choice helps to gauge where policies sit on a spectrum of vague to specific. Scanning the policies for keywords in LIWC's *Certain* and *Tentative* sections proxies lucidity. Pictured here are word clouds of those results for quick digestion. Both Spotify and Apple Music frequently use the words *may*, *any*, and *some*. Apple Music writes *approximately* four times, while Spotify does not once.

Spotify



Apple Music

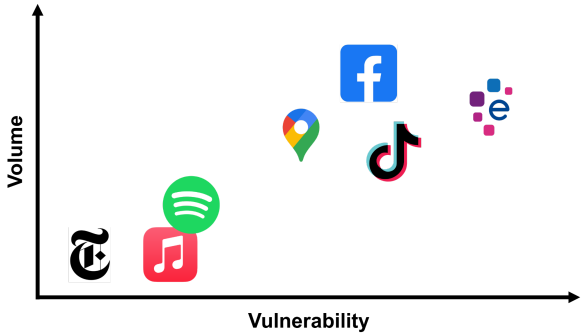
Spotify and Apple Music take noticeably different approaches to their privacy policies. Spotify is verbose where Apple Music is reserved. Spotify takes steps to incorporate friendly design (such as with formatted tables) to encourage comprehension while Apple Music conceals disclosure in paragraphs.



Spotify makes a clear effort to be more detailed about what personal data it collects, how it utilizes information, and what it does with it after processing. For these reasons, Spotify gets a higher Lucidity score than Apple Music.

Sensitivity

Spotify and Apple Music have roughly equal sensitivity, operating as more or less the same product in the same industry. Neither collects highly sensitive data when you consider the context of what other products and other industries collect. Social media apps like Facebook and TikTok, for instance, collect *much* more vulnerable information, as do data brokers like Experian.

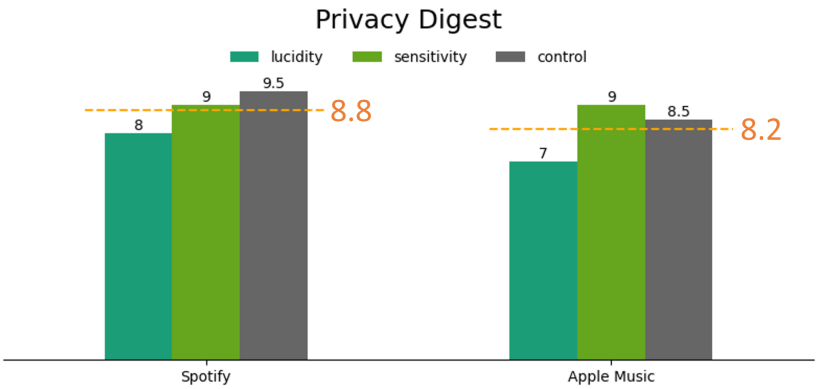


Control

Spotify provides considerable control to users, with privacy settings easy to find and digest. Apple Music does not provide standalone privacy settings for the music streaming product. Rather, it wraps these settings up in the same dashboards for iPhone, Macs, and other Apple products. This reduces a user's ability to opt-out, since they may have differing needs and contexts for their different services. For example, they may very well want location-services enabled on their mobile phone, but not want to allow that data to be used for playlist curation. Spotify's giving ample control over the use and share of data gives it a higher score than Apple Music.

Overall Scores

Spotify scores more favorably than Apple Music, primarily because of its privacy policy's specificity and easy access to opt-out controls.



Benefits of the Approach

Privacy Digest delivers three key advantages to advance data privacy:

1. It subverts dataveillance.
2. It leverages privacy expertise.
3. It may pressure organizations to be more transparent.

Subverts Dataveillance

Privacy Digest helps to subvert dataveillance by giving people an informed opportunity to opt-out. They may see a low score, click to learn more, and realize they aren't willing to trade location tracking for music streaming. And it meets users where they are. They do not need to navigate to a rankings-website each time they want to click an outlink, rather, the modal proactively embeds adjacent to that outlink. This ease is key to proactive intervention soliciting users to consider privacy in their decision-making

The approach can be classified as a combination of *discovery*, *avoidance*, and *refusal moves*, as defined by Professor Gary Marx in his book about neutralizing and resisting surveillance. Discovery moves are used to "find out if surveillance is in operation and where it is." Avoidance moves use a "temporal, geographical or methodological displacement to times, places and means in which the identified surveillance is presumed to be absent or irrelevant." And refusal moves "*just say no* and ignore the surveillance." (Marx 2003).

Privacy Digest discovers when it proactively identifies what data the site collects and how it utilizes that information; it *avoids* when users decide not to click through because they were deterred by a score or the detailed explanation; and it *refuses* when users change their privacy settings to opt-out of collection methods.

Leverages Expertise

The approach improves comprehension. Privacy policies are verbose legal documents that require nuanced understanding to truly grasp. As it stands today, web users have little chance of

protecting themselves, even if they are cautious technology experts. *Privacy Digest* makes it possible to proactively make decisions about *what* data to offer, *when* to do so, and *whether* to opt out, all by leveraging the expertise of professional privacy analysts.

Pressures Organizations

In addition to the primary goal of returning agency to individuals, a secondary benefit is rebalancing the power dynamics between data collectors and consumers. With sufficient traction, *Privacy Digest* could pressure organizations to be more specific and restrained in their privacy policies. Surfacing unethical policies, reporting on repeated infractions, and public pressure could coerce organizations to be more transparent or even change problematic policies. Giving individuals clarity in the trade-off between competitors may incentivize them to choose the option that is better for their long-term data health, even if that means incrementally higher subscription costs and considering alternatives to mainstream providers.

Limitations and Risks

There are at least four limitations and risks to this product concept:

1. Scores are approximations of risk.
2. Scores need time to calibrate.
3. Organizations may object.
4. User adoption is difficult.

Scores are Approximations

The clear trade-off here is nuance for simplicity. *Privacy Digest* is a score, an approximation of risk for individual users. The scores proxy overall risk for an average user. Some users will experience harm from sites with favorable scores on *Privacy Digest* and other users will escape unscathed from unfavorable scoring sites. The challenge of proxying is particularly pertinent to the Lucidity score, which is more qualitative than quantitative. A statement that reads as vague to the average user may be meaningful to a privacy expert, just as users with product familiarity may find nuances that newcomers overlook.

Judgment Subjectivity

When *Privacy Digest* employs analysts to assign scores, it adopts not only the benefits of human judgment, but also bears the downsides of subjectivity. Analysts may be inconsistent between one another; one analyst might score a policy's Control favorably where another might find it mediocre. And analysts may exhibit inconsistency versus themself; on Monday an analyst might be more critical and score sites generally lower than they do on Tuesday. Renowned psychologist Daniel Kahneman covers this challenge thoroughly in his books *Thinking Fast and Slow* and *Noise* — human judgment is less consistent than we wish it to be. Judgment deviates with a multitude of measurable and invisible variables including whether we've had a good day and if we have emotional attachment to a consideration. And this judgment noise is just the start, we all have biases as well. It is worth considering whether *Privacy Digest* would better be implemented as a checklist evaluation of what a policy does and does not contain, which is more explicit, relying less on subjectivity.

Calibration Lag

Scores will calibrate over time. The design intends to take into account ongoing developments, such as changes to privacy policies and reports of infractions. In this way, the scores should converge to a theoretical "truth" over time and gain greater-and-greater utility.

Legal Action and Disputes

Privacy Digest sets out to benefit individual web browsers, not web hosts or the parent organizations that own them. Organizations may disagree with their scores or characterizations of their policies. These organizations may attempt to combat unfavorable scores by pursuing legal action against *Privacy Digest* or taking action to diminish the product's reputation.

User Adoption

Like any new software or method, *Privacy Digest* needs traction before it returns agency to individuals or exerts positive pressure on organizations. Asking individuals to install *yet another* extension may meet resistance. A partnership with web browser providers that pre-installs the offering for users could be one tactic to market penetration. It will take consistent accuracy, marketing to demonstrate successes, to get users to trust the scores. Ultimately this is why the scores must be kept straightforward and stand in contrast to the opaque policies they rate.

Website Security

As of December 2023, *Privacy Digest* scores do not account for data security. It could be worthwhile to consider an additional score that would evaluate the website's security measures, such as whether the site uses HTTPS protocols, requires multi-factor authentication, and has stated policies for actions to prevent data loss.

Scraping Practicality

A practical limitation that could slow the development of *Privacy Digest* is whether web-scraping of policies is more intricate than expected. Websites that have technical safeguards in place to prevent scraping will slow the data acquisition portion of the process down. Similarly, websites that distribute their data privacy policies throughout several web pages, rather than consolidated to a single page, may present a need to have analysts comprehensively peruse entire websites to ensure all relevant pages are accessed.

Conclusion

Data collection and brokerage continues to accelerate. It's fair to consider that the nightmare privacy scenario Rita Raley described in 2013 could become reality unless we collectively turn the tables. "Perfect anonymity is impossible, but the nightmare scenario (then and now) imagines a womb-to-tomb "record prison" or "database of ruin," a massive "database in the sky" held by Google or elsewhere that contains the material necessary to reduce the entropic uncertainty about individual identities and thus cause demonstrable and legally recognized harm to everyone recorded within it." (Raley 2013). Part of the change we need could look like

Privacy Digest, with data and tools to empower individuals combined with legislation and market forces exerting influence for good corporate behavior from the top-down.

References

1. Apple Music & Privacy. https://www.apple.com/legal/privacy/data/en/apple-music/. Accessed November 26, 2023.
2. Apple Music privacy settings. https://www.apple.com/privacy/control/. Accessed November 26, 2023.
3. Auxier, Brooke; Rainie, Lee; Anderson, Monica; Perrin, Andrew; Kumar, Madhu; Turner, Erica. *Americans' attitudes and experiences with privacy policies and laws*. Pew Research Center; 2019.
4. LIWC. Linguistic Inquiry and Word Count website. https://www.liwc.app/. Accessed November 29, 2023.
5. Martin, Kelly D.; Borah, Abhishek; Palmatier, Robert W. *Research: A Strong Privacy Policy Can Save Your Company Millions*. Harvard Business Review; 2018.
6. Marx, Gary T. *A Tack in the Shoe: Neutralizing and Resisting the New Surveillance*. Journal of Social Issues; 2003.
7. Muthukadan, Baiju. *Selenium for Python*. Selenium Documentation; 2018
8. NewsGuard website. https://www.newsguardtech.com/. Accessed November 28, 2023.
9. Nissenbaum, Helen. *A Contextual Approach to Privacy Online*. Daedelus; 2011
10. Privacy Policy. Spotify Technology S.A. website. https://www.spotify.com/us/legal/privacy-policy/. Accessed November 26, 2023.
11. Raley, Rita. *Dataveillance and Countervailance: Chapter 7, Raw Data is an Oxymoron*. The MIT Press; 2013.
12. Regular expression operations in Python website. https://docs.python.org/3/library/re.html. Accessed November 28, 2023.
13. Solove, Daniel J. *A Taxonomy of Privacy*. University of Pennsylvania Law Review; 2006
14. Spotify privacy settings. https://www.spotify.com/us/account/privacy/. Accessed November 26, 2023.
15. Wagner, Isabel. *Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996--2021*. Cornell University; 2022.

Appendix

NewsGuard Acknowledgment

The human reviewers supported by data approach and delivery mechanisms are heavily inspired by NewsGuard, a rating system for news and information websites that supports browser extensions and plug-ins to commonly used tools such as Slack (NewsGuard 2023).

Positionality & Reflexivity

I am technology-oriented and believe in the power of data to transform businesses and behaviors. This predisposition may make me overconfident that a technical solution is the answer to returning data protection agency to individuals. It may well be that the more holistic approach is to pursue regulation and pressure web providers top-down to encourage users to read privacy policies proactively. My positionality may lead to an over-engineered solution that fails to appreciate human behaviors and alternative approaches.

I am a well-compensated worker and student with copious bandwidth and discretionary funds to spend time researching privacy policies and picking alternatives without consideration for cost. Not all individuals have those luxuries, and may choose free and low-cost providers even when it is worse for their long-term data health. My positionality may lend me to build solutions that assume users have time and assets to make tradeoffs.

I am an active user of Spotify and vocally favor it over Apple Music and other music streaming services. While the two products are included in this report purely as examples, my rating of Spotify's privacy policy may be influenced by my personal experience of the service.

Links

Slides: https://drive.google.com/file/d/1GZhOMdibeGN4iTPngabBm-pQG5MZVYYO
GitHub Repo: https://github.com/drussel4/Privacy-Digest