Responsible Data Collection Strategies for News Organizations

Written By Ben Werdmuller

Every reader interaction with your newsroom generates valuable data. A reader signs up for your breaking news alerts. Another downloads your voter guide. Someone else fills out a community survey about housing concerns.

Each interaction creates a trail of data that, used wisely, has the potential to help you serve your audience better. It can also breach their trust.

First-party data is strategically valuable because it connects users' preferences and behaviors to specific user profiles, often anchored to an email address. That's powerful – but it also presents a slew of ethical and legal considerations for every newsroom collecting it.

A Quick Guide to Audience Data Types

- **Third-party data** is collected by outside companies and shared with others, such as data aggregators, data brokers, or advertising platforms. It's broad, often aggregated, anonymous and outdated, and increasingly restricted by privacy regulations.
- **Second-party data** is collected directly from users by another organization and shared with you. It's someone else's first-party data that's shared in a partnership.
- **First-party data** is collected directly from users on owned channels, such as websites, apps, newsletters or events. It can

include both known and anonymous users, depending on the gating strategy of the publisher.
- **Zero-party data** is provided intentionally by a user to a company, like survey responses or preference settings. It's often considered a subset of first-party data, and also referred to as declarative data.

Credit: Kevin Charman-Anderson

To understand your data, you need to connect the dots.

In most cases, this data isn't stored in one place. Your newsletter subscriptions and activity logs are locked up in one platform, your website analytics in another (or, if you're using both Google Analytics and Parsely, more than one), your event registrations in another, and your social media activity in a few more.

This constellation of data sources makes it very hard to learn more about the complete journey of your audience as they discover your journalism, engage with it, and become a subscriber or donor. For example, you might discover that readers who engage with your housing coverage on social media, then visit your voter guide, are 3x more likely to become newsletter subscribers — but you'd never see this pattern with data trapped in separate platforms.

These platforms aren't designed to connect seamlessly. While basic tracking (UTM codes, browser cookies) helps, complete analysis requires extracting data into a central store via integrations. These can vary from simple Zapier workflows (no-code data transfers) to complex pipelines using custom code or platforms like Fivetran. Smaller newsrooms might

use Airtable as a data store; larger ones with more technical resources might choose BigQuery or Snowflake. These need to connect to dashboarding tools to help newsroom teams make sense of the information.

However, consolidating data across platforms introduces additional legal and security considerations that newsrooms should navigate carefully.

Many services have specific restrictions on data export, third-party sharing, or commercial use of exported data, particularly on free tiers. Your current service agreements may not permit the kind of data extraction and centralization you're planning, so audit these before building integrations.

You should also consider the security implications of your data pipeline. Your consolidated data is only as secure as the weakest link in your toolset. Evaluate whether each tool in your stack meets appropriate security standards. In particular, look for encryption at rest and in transit, robust access controls, and statements about which third parties might handle the data. For sensitive information like source communications or reader surveys about controversial topics, you may need to keep certain datasets separate rather than centralized.

Finally, ensure your privacy policy accurately reflects your data integration activities. Readers may have originally consented to data collection by specific platforms, not necessarily to cross-platform analysis that creates more detailed profiles. It’s a good idea to update your consent language to clearly explain any data consolidation before implementing it.

It may be worth starting with less sensitive data to test your processes. Begin with newsletter engagement metrics or website analytics before

moving to more personal information. Document your data flows and security measures throughout in order to help identify potential vulnerabilities before they become problems.

Data stewardship requires intention.

Once you have the technical infrastructure to collect and analyze data, the challenge is to use it in a way that upholds your commitments and responsibilities to your audience.

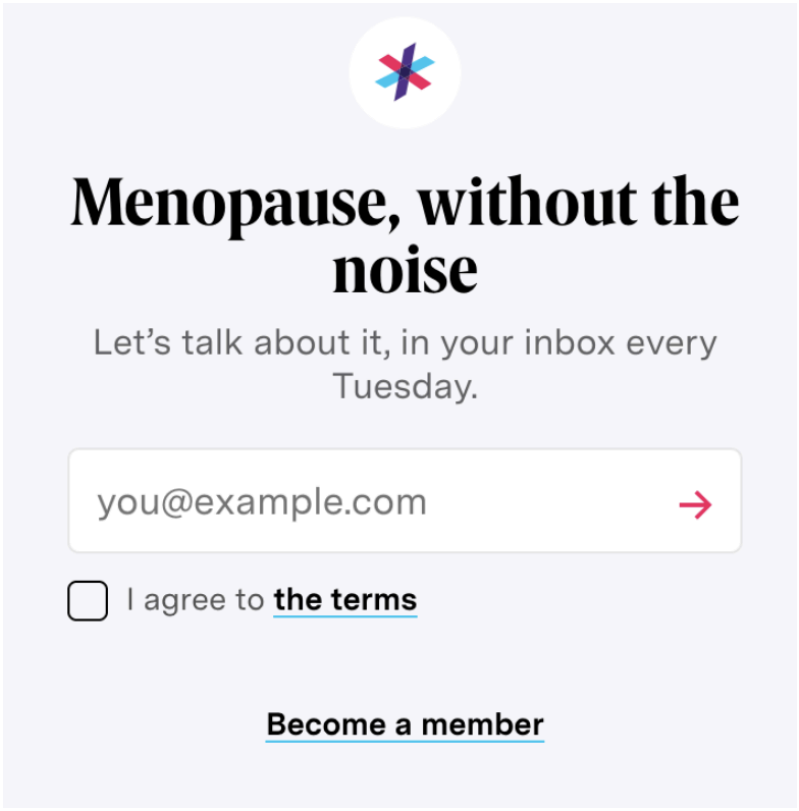
Every data point represents a real person who has chosen to engage with your organization. How you handle that information affects not just your ability to serve them better, but their willingness to trust you with it.

Consent is key. All organizations collecting user data should be transparent about their practices and give users meaningful control over their information. This means using clear language in privacy policies and making it straightforward for people to understand what's being collected and why.

With first-party data, this transparency becomes even more critical because you're directly asking people to trust your organization with their information, often including their email address and detailed behavioral patterns. Your audience deserves to know not just what data you're collecting, but how you plan to connect it across platforms. Add brief explanations on sign-up forms ("We'll use your email to send our weekly newsletter and analyze which stories resonate most"), and make it easy for readers to understand and control their data.

Consider opt-in rather than opt-out approaches for non-essential data collection, and make it easy for readers to request that their information is removed. You should describe the purpose the data is used for and

stick to it. If you haven't told them that quotes from their editorial survey responses might be included in fundraising materials, you shouldn't do that. That isn't to say that you can't use trends you identify in your fundraising – but you shouldn't use specific responses without permission.



The 19th's newsletter subscription form clearly gathers consent.

Collect with purpose. Any data collection should serve a specific, identified need. Organizations should be able to articulate which question each piece of data will help answer before collecting it. If you can't explain the purpose, it's better not to collect it in the first place. First-party data collection requires even more discipline because the data is directly tied to individual users who have chosen to trust you with their information. Unlike anonymous analytics, every data point you

collect can be connected back to a real person who chose to engage with your newsroom.

Every data point you store creates ongoing costs: security vulnerabilities that need protecting, compliance requirements that demand attention, and storage expenses that compound over time.

Irrelevant data also creates noise that obscures the insights you actually need, leading teams to chase false patterns or spend valuable time analyzing information that doesn't drive decisions. By collecting only what serves a clear purpose, you reduce these risks while ensuring your data remains a strategic asset rather than an operational burden.

Consider potential harm. All data collection carries potential risks if information falls into the wrong hands. Organizations should evaluate whether the data they're collecting could be used to cause harm to users if accessed by bad actors or authorities.

First-party data amplifies these risks because it creates detailed, personally identifiable profiles. **When readers sign up for your immigration newsletter or download your abortion rights guide, they're not just expressing interest—they're creating a documented connection between their identity and potentially sensitive topics.** You should consider whether the data you're collecting could be used to cause harm if it fell into the wrong hands.

For example, if you're collecting audience questions about abortion restrictions, you should avoid collecting personally identifiable information that could lead to legal consequences for users living in jurisdictions with restrictive reproductive laws. If you're collecting logged-in behavioral data that connects a user's identity with their reading history, consider whether you might accidentally implicate them.

If you have to collect that information in order to provide an essential service, ensure that your systems are up to it. Are you encrypting data at rest as well as in transit? Do the services you're using give themselves the right to transmit your data to third parties as part of their privacy policies? It may even be worth moving out of the cloud onto on-premise storage if you're collecting very sensitive data like immigration status.

Avoid unfounded assumptions. Data analysis should always distinguish between what the data shows and what it might imply. Organizations should be careful not to make assumptions that go beyond what their data actually demonstrates, particularly when those assumptions could affect how they serve different groups.

First-party data makes this distinction even more important because you can directly contact these individuals and your assumptions may shape how you communicate with them. Making incorrect assumptions about anonymous website visitors is problematic; making them about people whose email addresses you have can damage real relationships and trust.

There's a lot of talk about the value of contextual data – what you can infer about a person based on their behavior, such as what stories they click on most often or what newsletter they sign up for. You can reasonably assume that readers who engage with education stories are more interested in school board meeting coverage, but you don't know that they're parents unless they proactively tell you. Similarly, you can't know someone's ethnic background, gender identity, or political affiliation unless they tell you themselves.

If it's critical to your strategy to know people's identities, then you should explicitly ask rather than attempting to infer. Education Week asks readers to identify their role in the education space so that they can

do targeted outreach without making assumptions. But don't ask for more than you need.

Protect your sources. All organizations should consider whether their data collection practices could compromise the safety or privacy of vulnerable individuals who interact with them. This means evaluating whether the information being gathered could be used to identify or harm people if accessed by unauthorized parties.

For news organizations, this responsibility takes on special urgency when dealing with sources who provide information under the expectation of anonymity or confidentiality. Sometimes you need to refrain from gathering any data at all. If you're collecting tips on your website from anonymous sources, you should consider taking steps to ensure they can't be identified. That includes removing your website analytics from those pages and, if you can, preventing your web server from logging their IP address.

Even routine first-party data collection can compromise sources if you're not careful. A source who signs up for your newsletter about immigration enforcement, then later submits an anonymous tip, could potentially be identified through behavioral patterns or timing correlations in your data. Any time you store data in the cloud, it could potentially be obtained by law enforcement or third-party hackers. Even if the worst happens, you should honor the trust sources have placed in your newsroom by sending information to you. The Freedom of the Press Foundation provides good advice about source protection.

Staying compliant means understanding the principles behind regulatory changes.

The regulatory environment around data collection is evolving very quickly. Europe's GDPR was quickly followed by California's CCPA, and more legislation is emerging around the world. Some of these regulations have carve-outs for non-profits; others do not. It's undeniably complicated.

These laws generally have the following underlying intentions:

- Allow people to understand, correct, and delete the data that has been collected about them.
- Ensure data about a person cannot be secretly communicated to a third party for its own purposes (for example, to sell their information as part of a marketing list).
- Ensure individuals' data is stored securely.

You should always seek legal counsel before making any decisions for your organization, but these practices can keep you on the right track even as specific regulations continue to evolve:

Map your data collection. It's a good idea to maintain a list of every system that gathers data about your audience and how it is combined, transferred, and stored. That way you can easily provide it if asked for. This also allows you to maintain accurate privacy disclosures. If you don't have a list like this yet, you can use Kevin Charman-Anderson's instructions on how to conduct a data census to get started.

Be transparent about your data collection. Let readers know exactly how you're using data about them, and ensure you're not going beyond what you've declared to them. Ideally, let them opt in to specific uses. Consent management systems – which help organizations manage user consent data, typically for the purpose of compliance with privacy laws – are expensive. If you need to build this feature yourself, remember to

record *what* your reader consented to and *when*, so that you can provide this information on request.

Allow people to see, correct, and delete what you hold about them. If you don't have a portal that supports this feature, share publicly that this can be done by request via email.

Consider the legal basis for the data you collect. For each field: did the reader consent for you to gather this? Is it non-invasive data that a reasonable person might assume was being collected to support your newsroom's business? Did they sign a contract, for example when they subscribed? Is collecting the field a legal requirement in your region?

Above all else, get legal advice to ensure you're compliant with local regulations and that your privacy policy covers you.

With great power ...

When collected ethically and applied with intention, first-party data has the potential to help newsrooms to understand their audiences better and use that information to improve their journalism and build more sustainable revenue models.

But with this capability comes responsibility. Every email address in your newsletter database, every click tracked on your website, and every survey response represents a person who has chosen to engage with your journalism. It's also personal information that can be misused, unintentionally or maliciously.

How you handle that information determines whether readers will continue to trust you with it.