Data Sharing for Research Case Study: Meta

Executive Summary

The Future of Privacy Forum (FPF) analyzed a diverse sample of data-sharing partnerships between companies and academic researchers and produced a series of case studies distilling our findings. We learned that there is broad consensus regarding the potential benefits of industry/academic data-sharing partnerships, including the acceleration of socially beneficial research, enhanced reproducibility of research breakthroughs, and broader access to valuable data sets. At the same time, companies and academic researchers understand and take steps to mitigate risks - particularly ethical and data protection risks. Increasingly, stakeholders are identifying risks arising from re-identification threats or data breaches while acting to mitigate those risks through the use of Data Sharing Agreements (DSAs) and Privacy Enhancing Technologies (PETs).

FPF's analysis of corporate-academic data-sharing partnerships provides practical, evidence-based recommendations for companies and researchers who want to share data in an ethical, privacy-protective way. These case studies demonstrate that corporate-academic data-sharing partnerships offer compelling benefits to companies, research, and society. Risks exist, but effective mitigation strategies can reduce the likelihood of harm to individuals, communities, and society. For many organizations, data-sharing partnerships are transitioning from being considered an experimental business activity to an expected business competency. This trend is most pronounced among established firms; it is an opportunity for researchers to access new data for scientific discovery.

Data Sharing Type

Closed Trusted Partnerships, Open Data

Organization and Partners

Company

Meta is a multinational technology company founded in 2004 and based in Menlo Park, California. Meta provides several platform-based services such as Facebook, Instagram, and WhatsApp, employs around 77,000 people, and reported an annual revenue of $116 billion in 2022.

Partnership Considerations

Data Sharing

Representatives from Meta stated that their approach to research data sharing has evolved over the last ten years. Product teams and cross-functional teams (legal, policy, academic partnerships, etc.) work together to enable data sharing. They communicated that there are four main stages for data sharing; 1. identifying researcher needs, 2. understanding how to ensure user privacy and data security, 3. building data sets, and 4. maintaining data sets. By starting with identifying researcher needs, they say they try to efficiently meet those needs while building something of value for the research community. Additionally, their work centers on user privacy while attempting to identify interesting data sets or increase data utility.

The team remarked on misconceptions that sharing data is easy, explaining that building data sets for sharing is a fairly complex process. They added that it isn't as simple as just running an SQL query to produce a data set ready to be shared. Oftentimes they have to combine data sets in specific ways to pass internal quality assurance requirements, and each process usually involves new work. If the team determines that the data they created is of sufficient quality and accuracy that it is fit for research purposes, they can

begin onboarding researchers to test and iterate the data as needed and confirm that it is fit for purpose. Maintenance of shared data requires different levels of support based on the researcher's needs. For example, if the data needs infrequent updates, the time required is less arduous. However, if the data needs to be dynamic or real-time, the time and effort requirements are typically much larger. In both cases, however, the team has to be available to operationally support the datasets and tooling.

Data Sharing Agreement

Meta representatives described the use of multiple forms of Data Sharing Agreements (DSAs) depending on the type of partnership being considered. They work with researchers' institutions to ensure DSAs meet the needs of everyone involved. Meta leveraged Social Science One in its effort to negotiate a standard DSA for researchers to request Facebook data for certain research questions. The data-sharing team expressed support for the European Digital Media Observatory's (EDMO) working group's approach to data-sharing agreements. Additionally, the Inter-university Consortium for Political and Social Research (ICPSR) agreed to host data from Facebook and Instagram related to the US 2020 Election and has its own DSA to which researchers requesting access to data must agree. Their DSAs also address scientific oversight, an area where 3rd parties can be useful. If researchers want to use sensitive data in a publication, Meta can stipulate that it can review the data prior to publication to ensure user privacy isn't compromised.

Data Sharing Frequency

Representatives communicated that they regularly engage in data sharing with researchers, but the frequency depends on the project. For example, their Meta ads library, a dataset of all the ads running across all Meta products that do not involve personal data, is offered 24/7 via an API, and an ad will appear in the Ad Library within 24 hours from the time it gets its first impression. Any changes or updates made to an ad will also be reflected in the ad library within 24 hours. More focused data-sharing partnerships may involve fewer steps or deliverables, so the frequency of data-sharing can change

depending on how it's defined. The team commented that 'the right amount' of data sharing is a moving target. The resources that the company dedicates to data sharing, such as staffing or funding, can change over time, which affects the capacity of data sharing they can engage in. The team added that they draw from guidance provided by both EDMO and FPF's Playbook: Data Sharing for Research to help inform when to make data readily available for researchers and what mechanism to use for sharing.

Data Privacy and Sharing

Meta representatives said they conduct a privacy review for data proposed to be shared in a publication. The use of Privacy Enhancing Technologies (PETs), such as differential privacy, encryption, data aggregation, de-identification, or K-anonymization for data sharing depends on the project. Factors such as the sensitivity of the data and the mechanism for its sharing (direct transmission, researcher API, data clean room, 3rd party, etc.) all influence how privacy is approached. There is often a balancing test among data sensitivity, security, and utility when identifying the appropriate safety levels needed to share data. There are no hard requirements on what technology is used as there are a lot of moving parts for each partnership. Regardless of the technique used, the team considers how much data privacy protection is needed and how those techniques introduce bias and variance into the dataset. The team has to clearly communicate with researchers about the statistical and analytical impacts of privacy techniques so researchers can account for them in their analysis.

Costs

Meta's representatives added that their experience demonstrates how data sharing takes time, effort, and technical infrastructure, all of which translate into costs. The team expressed that, while a one-time data set release may be less expensive, it may also have less utility for research than a longitudinal dataset and that utility tradeoff should be balanced in terms of development cost and use of internal capacity. Additionally, any data-set release - one time or longitudinal - also needs to be balanced against developing tooling that enables access for researchers at scale. Researcher interest in longitudinal

data can lead to both massive quantities of data and added operations support. In the case of datasets that are so large they make data transfer impractical, further expenses such as hosting and computation are required.

Risks and Benefits

Risks

The data-sharing team said that the absence of clear regulation or codes of practice regarding things like liability structures and vetting and the responsibilities of researchers leave it up to companies to make many data-sharing decisions on their own. Meta attempts a risk-based approach that focuses on risks to users in choosing what data to share and how to share it. Supporting privacy-protective research also comes with reputational risks, especially if that research can be critical of the company that's sharing it – a salient risk for platform businesses today. There's also a concern about the potential misuse of data by researchers. In Meta's DSA with Social Science One, the company's agreement is with the academic institutions as co-signatories with the researchers. Platforms put a lot of trust in academic research institutions, which the DSA codifies. Researchers affiliated with universities have their own ethical codes of conduct and review boards, which operate as additional safeguards, and universities are long-lived legal entities that can take on liability, all of which contribute to risk mitigation. Meta is interested in how data-sharing governance structures on the company side interact with data-sharing governance structures on the research side, in particular, how they can work together to reduce data-sharing risks for everyone.

Benefits

Data sharing as an activity has allowed Meta to learn a lot, both about the findings of the research produced as a result of sharing, and about the processes required to support it. They described data sharing is an act of scaling research. They pointed to the Data for Good program and the Social Capital Atlas as demonstrations of the social benefit that data sharing for research can provide. Programs like this can inform data-driven policy,

improve urban planning, and generally be used to inform the public. Meta flagged exemplary research that leveraged its data to generate valuable insights, such as the equity-focused work of Raj Chetty, as an illustration of the societal benefit of its data sharing for research. It also remarked on its sharing of data with a third party, ICPSR, for use in analyzing the role of platforms in the 2020 election

Partnership Information

Meta: https://about.meta.com/
Meta- Illustrative list of publications from data-sharing partnerships: https://developers.facebook.com/docs/url-shares-dataset/featured-works
Meta- Data for Good: https://dataforgood.facebook.com/
Meta - CrowdTangle Data for Researchers: https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers
US 2020 Election Project: https://research.facebook.com/2020-election-research/

To learn more about data-sharing partnerships, read The Playbook: Data Sharing for Research or join the Ethics and Data in Research Working Group for updates on legislative developments and monthly calls with experts. This project is supported by the Alfred P. Sloan Foundation, a not-for-profit grantmaking institution whose mission is to enhance the welfare of all through the advancement of scientific knowledge.