Data Sharing for Research Case Study: Microsoft

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

Organization and Partners

Company

Microsoft Corporation is a multinational technology company founded in 1975. In 2022, it reported $198 billion in annual revenue and employs roughly 221,000 people worldwide.[1]

Finding Partners

Microsoft would like to increase its data sharing, especially around programs for social good. Company representatives communicated that it can be hard to match data with researchers outside of Microsoft Research. Microsoft co-founded the Industry Data for Society Partnership to help overcome the fragmented nature of the data-sharing ecosystem. Microsoft has found that it can generally get more traction forming partnerships when the projects and data are for general social benefit instead of those solely for economic goals or applied to narrowly defined sectors.

The company pointed to one example of data sharing for social good which, as part of their broader Airband Initiative, Microsoft publicly shared data about broadband usage and speed across the US so that researchers could investigate issues relating to closing the rural broadband gap. This project evolved during the COVID lockdown in 2020, which featured massive transitions to remote work, school, and healthcare. As the data was segmented by zip code, there was the potential for reidentification in rural areas. To mitigate this risk, Microsoft employed differential privacy techniques, such as adding statistical noise to zip codes with small populations.

A second example the company mentioned of a data-sharing partnership for social benefit is Microsoft's work with Answer ALS, a non-profit organization dedicated to curing amyotrophic lateral sclerosis (ALS), a neurodegenerative disease. This project allows

[1] Microsoft. 2022. Annual Report 2022. https://www.microsoft.com/investor/reports/ar22/index.html

patients with ALS to share their personal health information and data about their disease with medical researchers. As the project began, Answer ALS obtained patient-level consent before any data collection or sharing took place. Microsoft executives commented that privacy issues with data sharing are easier to resolve by planning for them before the project starts instead of trying to share existing data and implementing privacy protections retroactively. They added that technologically-based privacy controls aren't sufficient; they need to be used in concert with thoughtful data collection programs and appropriate administrative and social controls.

A third partnership example the company shared was with the United Nations' International Organization for Migration (UNIOM) on human trafficking. Data about trafficking victims and case records are extremely sensitive and high risk. To ensure the protection of privacy and safety of victims and survivors, Microsoft researchers used differential privacy techniques to create a synthetic public dataset that described victim-perpetrator relations. No person's specific identity or information was ever released, but research could still be conducted that helped counter human trafficking.

Microsoft representatives have found that data-sharing projects that align with environmental sustainability, accessibility, and health, in particular, help create momentum in forming external partnerships. When Microsoft engages with other companies about sharing their data, a common concern is that companies first assume they're being asked to open all their data to everyone. Clarifying expectations on the scope of the data-sharing partnership, establishing a commitment to share data only with appropriate privacy safeguards, and aligning with environmental, social, and governance (ESG) values facilitates more productive conversations. Additionally, Microsoft representatives communicated that data-sharing partnerships can benefit both ESG goals and create business value through innovations, such as enhancing internal decision-making processes and performance, as well as creating value-added services or products.

Partnership Considerations

Data Sharing Processes

Microsoft representatives reported they have multiple approaches to data sharing. For example, Microsoft Research Open Data freely shares non-sensitive data and is tailored for research, as is Microsoft Data for Society. Microsoft's social media subsidiary LinkedIn has a broad data-sharing partnership with the World Bank and focused arrangements with academic researchers and publications of its own analyses, which are sometimes used in research. Microsoft's subsidiary, GitHub, has its own program, too. This means that data sharing isn't a uniform pipeline or process across the company but often develops organically based on the various needs of the business, partnerships, or research. Microsoft's general approach to data sharing is to make data as open as possible, especially when that data or project is related to positive social impact, such as the Data for Society resource center.

Data Sharing Agreements

There is no standard data sharing agreement (DSA) across Microsoft due to the variety of partners, uses, and data sensitivities. Almost every external partnership has a different DSA. However, there have been some efforts to standardize DSAs using the Linux Foundation's Community Data Licensing Agreement 2.0. Company representatives would prefer a standardized DSA to increase the ease and pace of collaboration. Progress toward that goal has been slow due to the complexity and variety of data-sharing efforts.

Data Sharing and Privacy

Microsoft representatives explained that it is committed to protecting individuals' privacy in any data-sharing collaborations that involve personally identifiable information. Furthermore, some technologies, such as confidential computing, enable insights to be drawn from data without the data itself being shared. Dashboards and visualization tools are other ways of making data accessible rather than granting direct access to data sets. A full-spectrum approach to data sharing that includes everything from fully open to fully

closed data sharing leads to more collaborations. According to company representatives, exclusively considering open data risks losing out on potential partnerships with people willing to collaborate using other kinds of data-sharing arrangements.

Costs

Costs for running data-sharing programs can include the time of key personnel, IT support, legal teams, data storage, communication, and computation, among others. Some projects can offer an economy of scale where particular costs go down, but this is not often the case. Egress fees for moving data from server to server can be a limiting factor. Representatives advised that planned data storage and transfer are two areas where standardized DSAs could help streamline data-sharing processes and reduce future costs.

Risks and Benefits

Risks

Microsoft representatives identified several risks inherent in data sharing. Historical incidents, such as in 2006 when AOL shared its users' search history with in-house researchers who were able to re-identify individuals, highlight the potential for severe consequences and discourage data sharing. Evolving a company's culture around data sharing is key. For example, complying with the General Data Protection Regulation (GDPR) can coexist with open data and data-sharing projects. These efforts can simultaneously account for privacy, security, compliance, and data utility.

According to Microsoft representatives, some of the risks for data sharing are perception-based and can be managed. They believe that once there are more good examples of company and social benefits to follow, more people will start overcoming the perceived risks and share data more often. There also needs to be community practices and norms for people to model. They referenced a quote from The Governance Lab at New York University that describes data sharing as "preventing missed uses of the data for solving

public problems"[2]. For Microsoft, this quote reflects a needed cultural change from legal and compliance-oriented fears about data sharing to a benefits-oriented assessment highlighting the missed opportunity to solve societal challenges if data isn't openly shared. By reframing the location of risk, or at least reframing where the emphasis of risk is, they believe more people will share data.

Benefits

Company representatives said they believe everyone can benefit from opening, sharing, and collaborating around data to make better decisions, improve efficiency, and tackle some of the world's most pressing societal challenges. They also stated that being more open with data can lead to more value derived from that data versus keeping the data siloed. Representatives noted that external stakeholders are often surprised when they learn about Microsoft's open data initiatives and are interested to learn more. They added that data sharing has led to new external relationships, new ideas, and made several important contributions to research and society. They point to 'The 9Rs Framework' from The GovLab[3] as a comprehensive description of the many benefits of data sharing, which helps to make the business case for why more companies should engage in it.

Partnership Information

Microsoft: https://www.microsoft.com/

Industry Data for Society Partnership: https://www.industrydataforsociety.com/

Answer ALS: https://www.answerals.org/

United Nations International Organization for Migration: https://www.iom.int/

---

[2] Saxena, S., Zahuranec, A., Verhulst, S. 2021. A Curation of Tools for Promoting Effective Data Re-Use for Addressing Public Challenges. The GovLab. New York University. September 29, 2021. https://blog.thegovlab.org/post/a-curation-of-tools-for-re-use

[3] Moretti, L., Zahuranec, A., Verhulst, S. 2022. The 9Rs Framework: A Worksheet for Establishing the Business Case for Data Collaboration and Re-Using Data in the Public Interest. The GovLab. New York University. https://businesscase.opendatapolicylab.org/

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.