REPORT

Data Sharing for Research

A COMPENDIUM OF CASE STUDIES, ANALYSIS, AND RECOMMENDATIONS

AUGUST 2023

FUTURE OF PRIVACY FORUM

BY

Shea Swauger
Senior Researcher for Data Sharing and Ethics
The Future of Privacy Forum

Marjory S. Blumenthal
Senior Fellow
The Future of Privacy Forum

WITH CONTRIBUTIONS FROM

Nicole De Brigard
Policy Intern
The Future of Privacy Forum

Jacob Leibowitz
Policy Intern
The Future of Privacy Forum

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.

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

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.

SIGNIFICANT FINDINGS

1. Companies discovered new ideas and improvements to their core products and services due to data sharing, often leading to reduced costs or increased value.
2. Public data sharing menus that identify and describe data a company is willing to share can facilitate data sharing partnerships.
3. There is a potential knowledge and infrastructure gap between companies and researchers regarding Privacy Enhancing Technologies (PETs) that creates barriers to data sharing.
4. Data sharing agreements (DSAs) continue to be an essential practice for sharing data, but a DSA's level of flexibility was newly correlated with company costs; the more flexible the DSA, the more expensive the partnership.
5. Different data sharing partnership modalities offer different affordances for privacy, risk mitigation, social impact, and costs.
6. Companies should approach data sharing like a product and allocate personnel, infrastructure, marketing, and institutional support accordingly.
7. In every company interviewed, the benefits of data sharing outweighed the potential risks.
8. Until there is more open data for researchers to work with, corporate data sharing offers the best chance for researchers to investigate new data, publish their results, and advance their discipline.

RECOMMENDATIONS

For companies that share data for research or are considering a data sharing program:

Transparency and Openness

- Create a public page that lists what data the company is willing to share, describe the data as much as possible, and update the list regularly. This could be done unilaterally or as part of a consortium of companies seeking to share more data for research.
- Be transparent about requirements that academic partners must meet and publicly post a typical data sharing agreement, if possible.
- Create a public form for researchers to ask questions, request data, or initiate a partnership. This signals the kind of information the company needs about the researcher and the proposed research.

Privacy and Security

- Use Privacy Enhancing Technologies (PETs) to bolster data privacy, but select PETs judiciously so as not to exclude researchers from less-resourced institutions.
- When using PETs is impractical, reduce the data's sensitivity (through minimization, aggregation, and other techniques) to a level that enables sharing while maintaining privacy.
- Ensure cybersecurity and privacy teams co-design privacy safeguards when sharing data.
- Include metadata as part of internal privacy reviews before sharing.

Governance

- Assign multiple people with expertise in data science, statistics, research, policy, and regulatory compliance to manage data sharing activities. This role is likely to be an extension of existing responsibilities, although, in some instances, a dedicated team might be feasible.
- Connect the general counsel, marketing/communications, and core software engineers to the data sharing team for effective coordination.
- Adapt the data sharing agreement to the amount of money and personnel available; more adaptable DSAs require more people and time.

Control

- Choose the appropriate data sharing partnership type: open data, closed trusted partnerships, and data intermediaries all require investment, personnel, and institutional support but can vary in duration and intensity.

- Companies should consider implementing a spectrum of data sharing models (open data to closed trusted partnerships), which would likely lead to more collaborations and greater social impact.
- Ensure researchers have full authorial control over the publication venue and all content not directly relevant to the data.
- When appropriate, reserve the right to review data before publication to assess privacy risks, accuracy, or any analytical limitations of the data.

For researchers interested in using data held by a company for research:

Initiative

- If a company doesn't have a data sharing menu or public page describing its data sharing activities, contact them to inquire about a potential partnership.
- Provide lots of communication on the front end of the partnership and plan check-ins at key points of the research lifecycle through publication.

Internal Partnerships

- When starting a data sharing partnership, involve the university general counsel early on and check to see if the university has a standard data sharing agreement or an example agreement used for a previous project.
- Contact the university's Research Integrity Office and Information Technology Office before any data is shared to ensure the university can support the project technologically and ensure regulatory compliance.
- Contact the university library for additional support for research, data management, and privacy.

Privacy and Security

- Get training on Privacy Enhancing Technologies and contact all relevant technical support offices to ensure the university can support data sharing using PETs.
- Include data transfer and privacy-related technical infrastructure in all research funding proposals or projects.

Communication

- Coordinate with the company about requirements for publishing, data sharing, data retention or deletion, citation, and promotion of the research to maximize the audience and impact.
- Maintain academic independence of any research produced from the data sharing partnership, but, when appropriate, allow companies to review data to assess privacy risks, accuracy, or any analytical limitations of the data before publishing.

TABLE OF CONTENTS

TABLE OF CONTENTS

INTRODUCTION

Corporate-academic partnerships to share data for research have the potential to benefit science, industry, and the public significantly. To accomplish this, companies must share data that might contain highly sensitive information with external researchers while protecting the identities and privacy of the people the data is about or from. This challenging endeavor introduces privacy, legal, and reputational risks for all parties involved. Fortunately, most risks can be mitigated with the thoughtful implementation of technological, physical, and administrative safeguards, and the resulting benefits justify the venture.

Companies increasingly collect data through the ordinary course of business. Such operational data can be referred to as "administrative data," a category that first emerged for data associated with government activities.[1] Researchers in a growing variety of fields noticed the growth of administrative data and sought access for their research purposes. FPF published a report in 2017 to address this trend titled Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers.[2]

Data collected by companies, especially through social media, recently attracted the attention of the public, researchers, and policymakers alike. In 2021–2022, United States congressional representatives brought forward four bills that specifically address researcher access to data: the Platform Accountability and Transparency Act, the Social Media Data Act, the Digital Services Oversight and Safety Act, and the Kids Online Safety Act. As of this writing, ten comprehensive state privacy laws contain exceptions to facilitate the voluntary sharing of data with researchers, and on November 16, 2022, the European Commission's Digital Services Act entered into force, requiring some online services to provide data to researchers. The legislative landscape is giving every indication that, in the near future, many companies may be obligated to routinely share data with researchers and create the necessary policies and procedures to meet that obligation without jeopardizing user privacy.

The Future of Privacy Forum (FPF) recognizes the power and potential available in research and has invested in identifying and uplifting privacy-protective best practices for companies and researchers to utilize. In 2020, FPF began celebrating success stories involving corporate data sharing for research through an annual Award for Research Data Stewardship, casting a wide net for nominations and generating publicity for the award ceremony and the awardees. The third annual award was held on May 10th, 2023, and the winners demonstrated the scale of organizational commitment to privacy-oriented data sharing for research. In 2022, FPF synthesized guidance from extensive conversations with stakeholders from companies, researchers, and institutions to create The Playbook: Data Sharing for Research.[3] At the same time, FPF began convening and expanding those stakeholders to form the Ethics and Data in Research Working Group that hosts monthly meetings and offers legislative updates and analyses related to data sharing.

Although the award and working group are expected to continue, this report is the culmination of a period of intensive exploration and documentation of data sharing for research activities. This report presents a series of brief case studies of companies that share data with researchers. Whereas earlier FPF research products focused on enumerating and synthesizing key data sharing issues, this report aims to provide more detail on specific strategies and tactics companies and researchers can use to engage in successful corporate-academic data sharing partnerships.

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.

TYPES OF DATA SHARING

FPF researchers intended to produce a taxonomy of the different types of data sharing for research that companies use. However, classifying data sharing types proved more difficult than expected, which FPF considers significant in itself. While having an easily-referenceable chart that identifies all possible data sharing types and explains their similarities and differences sounds like a helpful resource, the interviews with companies and researchers proved that data sharing is more complex and situationally-dependent than can be represented by a taxonomy. When FPF researchers presented companies with draft taxonomies, every company responded that there were too many exceptions and conditional variables for the chart to accurately represent their data sharing activities. As a result, this report offers a high-level presentation of the four general data sharing categories we encountered in the interviews: internal data sharing, closed trusted partnerships, data intermediaries, and open data.

Internal: policies and procedures a company creates to regulate the flow of data from one internal sector to another for research purposes, often involving a review process that approves or denies sharing requests. This type of sharing is used when the data contains private information and/or is legally regulated and needs to be handled in specific ways. Sensitive internal data sharing that may trigger an internal review could include proprietary data or data with additional ethical considerations or that was collected under a narrow informed consent. Internal data sharing is the most restrictive form of data sharing addressed in the case studies.

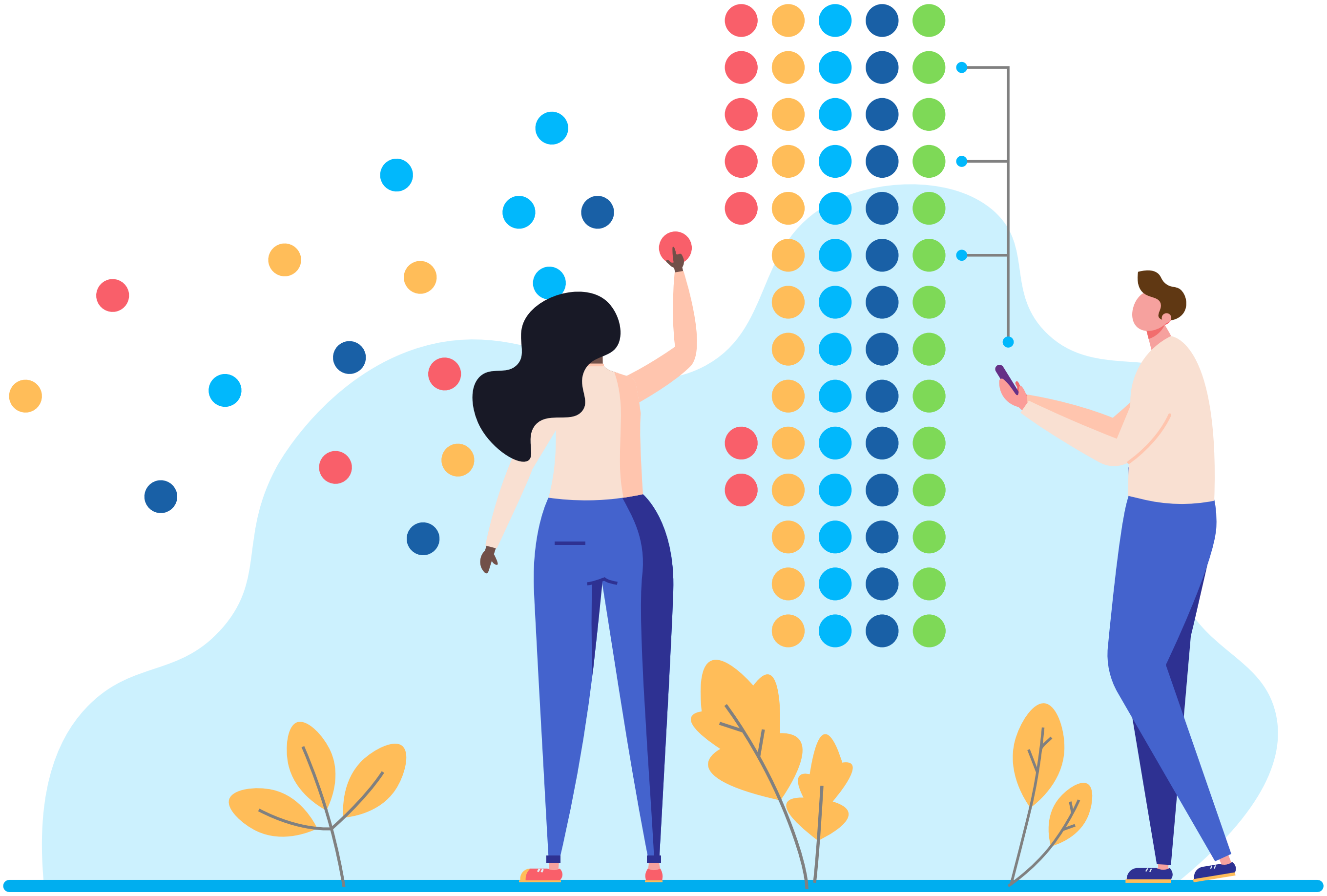
Closed Trusted Partnership: a relationship between a company and an external individual, group, or institution where the company shares data with external partners for research and under contractual obligations that address privacy and further data sharing. The partnerships are closed in that they restrict data sharing to within a defined population, and they are trust-based in that, while all data sharing expectations are codified through a contract, the parties involved consider them to be in a partnership that requires establishing mutual credibility before any data is transferred. Closed trusted partnerships are less restrictive than internal data sharing but more restrictive than open data.

Data Intermediary: an arrangement where a data creator (usually a company) transfers data to an independent entity that stewards the data and approves or denies access to it from external parties seeking to use it for research purposes. The term 'data intermediary' highlights the middle role of the partnership that connects the creator and end-user. Data intermediaries can vary in how strict or open their data access policies are. The YODA Project in the Johnson & Johnson case study is the only data intermediary in this report. FPF has previously called for increasing support for the data intermediary model for sharing administrative data for research.[4]

Open Data: a form of sharing data with little or no restrictions on who can access the data and what can be done with it. Open data is only appropriate for data that has no privacy risks or ethical sensitivities that might cause harm. Most of the data sharing for research reported in the interviews did not involve open data, with the exception of Meta and Microsoft (in some instances), but the interviews provided evidence that companies recognize the demand for open data, and some have taken steps toward more open data sharing.

Companies often reported using a combination of data sharing types. This allowed them to match the data it wanted to share with an appropriate data sharing type based on its sensitivity; open data for non-sensitive data, and closed trusted partnerships for sensitive data. Using multiple types of data sharing for research maximizes the potential for data use and subsequent social benefit.

DATA SHARING FOR RESEARCH CASE STUDIES



Presented Alphabetically

AIMS COLLABORATORY

Data Sharing Type
Closed Trusted Partnerships

Organization and Partners

Organization

The AIMS Collaboratory is a "community of practice with the goal of accelerating research and development in learning strategies for algebra education, especially for Black/Latinx students and students in poverty."[5] The community is composed of fourteen partnerships, and each partnership — called a trio — has three distinct members: 1) researchers; 2) schools or school districts; and 3) curriculum developers, some of which are educational-technology providers. For example, one trio called "Rice Algebra Initiative for Success and Equity (RAISE)" is a "trio" partnership between: 1) Rice University (researchers); 2) the Houston Independent School District, and 3) OpenStax (curriculum developer). Where common issues arise among multiple trios, cross-organizational efforts are organized to address them. Funded by the Bill and Melinda Gates Foundation,[6] AIMS has a facilitation team staffed by representatives from menloEDU, WestEd, and the National Network of Education Research-Practice Partnerships that organizes and supports the trios.

Partnership Considerations

Data Access

Data sharing among the trio members is an essential feature for the partnerships to be successful at improving math education, according to AIMS representatives, who added that while using a standard data sharing agreement for all partnerships would be desirable for efficiency, it has proved challenging. Most trios start with a standard data sharing agreement (DSA) template, and then individual members' legal teams usually add on specific terms and conditions to make the agreement appropriate for the trio. AIMS continues to adapt its standard DSA to make it useful for future trios with the goal of streamlining data sharing while protecting student data privacy and ensuring legal compliance.

Within each trio, data sharing can take different forms. Some trio members only share limited data with researchers, while others have fewer restrictions. In the trio between 1) the University of Toronto and the Abdul Latif Jameel Poverty Action Lab (primary researchers); 2) the Puerto Rico Department of Education (school district); and 3) Khan Academy (curriculum developer), some researchers are also employees of Khan Academy, which allows them greater access to data and more freedom to share within the trio. Increasingly, school districts are designating a single individual to manage external partnerships like these and support data sharing functions, with titles such as 'Director of Research' or 'Coordinator for Partnerships.'

While AIMS' mission is to "support education research," and data sharing is seen as essential to accomplish that under their model, the team cautioned that more data sharing isn't always better unless researchers employ principled and critical methods to avoid unanticipated and harmful consequences. For example, AIMS staff observed that some members of the data science community assume that if something shows up in data, it must be the truth, which is often not the case. They suggested that being able to validate data with members of the community can be an important quality check on data. They also remarked on a common tendency to collect all student data that can be collected rather than asking if it should be collected. These two behaviors, seeing data as objective truth and the proclivity to collect everything, are behaviors that AIMS has encountered across educational sectors, industries, and disciplines. AIMS staff reiterated that data should only be collected if it is directly tied to a research question and if there are adequate privacy and security protocols to keep data safe.

Ownership and Consent

According to experts from AIMS, questions about data ownership come up regularly in trios, sometimes with conflicting views among members that need to be addressed before the trio can move forward. For example, some members of a trio might argue that school districts legally own student data, but there are others who argue that the data belongs to the students and that the district is only a data steward. AIMS representatives did not express formal positions on what constitutes appropriate data ownership for trios but stated that competing data ownership approaches from trio members were a common obstacle. AIMS experts also identified consent as a regular issue that needs to be addressed when sharing student data. A common practice in AIMS is for school districts to get consent from parents and assent from students who are under 18. However, if a researcher or a school district is seeking data about students from a third-party platform, then they might not get parental consent if the data is generated through general classroom practices, such as a teacher assigning students to use the platform as part of instruction. According to AIMS, If third-party platform-generated student data is not collected as part of directed class time, then their process around legal ownership and consent may be different.

Costs

Data sharing partnerships in AIMS have a mix of start-up, ongoing, and ad hoc costs. Most trio members have legal teams in-house or on retainer that work on DSAs. The more standardized the DSA, the more efficient and cost-effective it may be for AIMS. A common cost question the organization faces is how to compensate school districts for staff efforts to perform data cleaning and minimization outside of their normal work. Otherwise, data cleaning and minimization can be done by a member of the research team or an external contractor, both of whom usually have to travel to the schools and do the work on-site under supervision. Grants might build in a stipend to compensate districts for data cleaning and sharing. Ongoing costs include data storage, transfer, security, and maintenance.

Risks and Benefits

AIMS identified several risks associated with data sharing partnerships like theirs, with the most significant being unintentionally exposing student data. There are secondary risks that follow something like a data leak or breach, such as losing the trust of students, parents, and community partners, or in extreme cases, the loss of professional reputation, grant funding, or employment. AIMS focused on two tenets identified as important for successful data sharing: first, thoughtful data privacy and security practices, and second, that there must be trust among all stakeholders. The parties recognize that all stakeholders have something to lose if sensitive data is handled improperly, especially students. Additionally, a single leak or breach of student data may incur a significant loss of trust and make future partnerships less likely, even if data privacy and security practices are improved after the fact.

The AIMS Collaboratory is built on the premise that data sharing can help to increase equity in math education. AIMS states that complex problems require a diversity of people and approaches to solve them, and their collaborations rely on ethical and privacy-oriented data sharing. The goal is to benefit students, the community, and improve K-12 math education. Additionally, AIMS hopes to support rigorous research methods and improve secondary data analysis practices, both of which will help close critical gaps to make education serve all students.

PARTNERSHIP INFORMATION

AIMS Collaboratory website: **www.aimscollaboratory.org**

J-PAL: **www.povertyactionlab.org**

University of Toronto: **www.utoronto.ca**

Khan Academy: **www.khanacademy.org**

Puerto Rico Department of Education: **de.pr.gov**

GRAVY ANALYTICS

Data Sharing Type
Closed Trusted Partnerships

Organization and Partners

Company

Founded in 2011, Gravy Analytics is a location technology company based in Virginia with about 60 employees and a reported annual revenue of $16.9 million in 2022.[7] The company primarily provides location data and analytics to other companies but also maintains a Data for Social Good program, which offers reduced rates for research-based institutions.

University of Florida

Dr. Xilei Zhao is an Assistant Professor in the Department of Civil & Coastal Engineering at the University of Florida. Dr. Zhao's research team maps the evacuation flows of people affected by natural disasters to inform public safety decisions. For example, the team explores how specific wildfires impact population movements, such as who chooses to evacuate, who doesn't, when, how they move, and why. Dr. Zhao partnered with Gravy Analytics on her team's research to map people's movement during the 2019 Kincade Fire in California and the 2021 Marshall Fire in Colorado. Their goal is to eventually be able to forecast and create real-time emergency management tools to inform responses to future disasters.

Columbia University

Dr. Sandra Matz formed a data sharing partnership with Gravy Analytics through the Data Science Institute at Columbia. Dr. Matz is an associate professor in the Business School at Columbia University. Trained as a computational social scientist, Dr. Matz studies people's online and offline behavior and the relationship between the events and locations they frequent. Dr. Matz wanted to be able to match an individual's psychological profile with the content they interact with online and where they physically spent their time, such as attending a concert or a political gathering. Her theory was that these factors can contribute to people's sense of identity, political ideology, and social values. Dr. Matz combined traditional psychology methods in controlled settings with Gravy Analytics data, allowing her research to investigate questions at a larger scale in applied settings and potentially infer causality.

Johns Hopkins University

Dr. Anton (Tony) Dahbura is the Co-Director of the Institute for Assured Autonomy and the Executive Director of the Information Security Institute at Johns Hopkins University. In 2020, as the COVID pandemic began, Dr. Dahbura and his team wanted to develop a community-level data center to provide accurate epidemiological models for how respiratory infections like COVID spread throughout a community and what kinds of interventions are most effective. Traditional epidemiology often uses agent-based modeling, where researchers create simulated people and give them activities, such as the need to find food or go to sleep, and then track how the simulated agents move around. However, agent-based models may not behave like real people within their communities, so Dr. Dahbura wanted to be able to model his simulation using real location data from Gravy Analytics to increase its accuracy.

Partnership Considerations

Data Access and Sharing

Researchers can apply to purchase data from Gravy Analytics at a reduced price through an online form on the Gravy Analytics' website. Company representatives said that applications are reviewed internally for feasibility, capacity, and mission alignment. Their application vetting process usually begins with a call with the requester to discuss their qualifications, institutional support, privacy considerations, the data requested, and research questions to be addressed. Gravy Analytics formalizes partnerships with a standard Data Services Agreement (DSA), which it can adapt based on the partner's needs. Representatives at Gravy Analytics indicated that private universities have demonstrated more flexibility in negotiating DSAs than larger public universities. In either case, Gravy depends on the institution for compliance with the agreement terms and data governance.

Each data request requires work from multiple teams, and Gravy Analytics estimates a single request can take up to 50 hours of internal work before anyone receives data. Data is shared only with the granularity required to answer research questions, which varies by use case. Gravy Analytics distributes data under a non-disclosure agreement (NDA), but, in concert with its DSA, it allows the researchers' results and findings to be shared and published. Every request requires data cleaning, organization, and the application of privacy techniques, as well as additional researcher obligations to protect data privacy and security.

University of Florida researchers identified several factors that went into their decision to partner with Gravy Analytics, including data accuracy and data use time limits. Gravy Analytics offered a three-year contract that met the needs of the team's research and grant timeline. Dr. Zhao recommended that researchers ask for sample data sets from potential data sharing partners. After comparing different companies, Dr. Zhao selected Gravy Analytics and signed an NDA for the data purchase. Dr. Zhao combined the purchased data from Gravy Analytics with other open-source data such as census, demographic, parcel, and land use data. All data purchased was funded by a NIST grant supporting the research project.

At Johns Hopkins University, Gravy Analytics provided Dr. Dahbura with about one month's worth of location data for the state of Oklahoma. The choice of state allowed the research team to demonstrate how COVID spreads in more isolated communities, as opposed to denser suburbs or large cities. Dr. Dahbura mentioned that the Institutional Review Board (IRB) process, data request, and transfer all went much faster than normal because the research was related to a public emergency (COVID), but generally, he felt that his research was well supported by the institution and legal team that worked to affect the data sharing partnership.

Privacy, Publishing, and Teaching

The University of Florida research team was given data that had already undergone aggregation and some anonymization from Gravy Analytics, but which still allowed them to answer their research questions. Despite these precautions, the research team decided that the shared data was sensitive enough to warrant storage in a secure environment, access restrictions, and several privacy techniques to minimize re-identification before publishing.

Dr. Matz noted that, in her experience with corporate data sharing, most companies expect a faster turnaround on the research produced from data sharing than academic publishing allows. Managing expectations with Gravy Analytics early on helped avoid conflicting expectations for project and publication timelines. Dr. Matz mentioned that corporate data sharing is largely impractical for students to engage with independently, as just getting DSA and IRB approvals can take over a year to obtain. Students often can't wait that long to access data; however, if they work with a faculty member involved in a long-term data sharing project, students can meaningfully participate.

Risks and Benefits

Risks

Company representatives said that location data is often very sensitive, so all data sharing must go through internal privacy and security checks to minimize risks to end users. Sharing location data with external partners may create reputational risks for Gravy Analytics and research partners, even when the underlying research is widely perceived as meritorious.

Columbia University offered Dr. Matz institutional support for the partnership with Gravy Analytics, involving the IRB as well as its legal team, which addressed ethical risks and legal obligations. While Dr. Matz uploads descriptions of the shared data, the methods used, and her code in an Open Science Framework instance, she explained that research reproducibility could be difficult with data sharing partnerships.[8] Many DSAs mandate data erasure after a certain period and prohibit data sharing outside the research team. Meanwhile, some academic publishers require that the data be kept by researchers for five years. Other researchers who want to reproduce her research based on the data sharing partnership could potentially petition Gravy Analytics for the data, but the decision to share it and at what price would be up to the company.

At Johns Hopkins University, Dr. Dahbura was concerned about the risks of handling sensitive location information, so the team worked with experts in cryptography and edge computing. They also processed the shared location data using AI to create a synthetic version of the data to model mobility patterns. This way, the model never directly used actual location data, significantly reducing re-identification risks. This method provides greater confidence in the simulation outcomes and increases the potential scale and demographic accuracy.

Benefits

Gravy Analytics concluded the benefits of partnering with research institutions outweigh the risks, including helping to legitimize socially beneficial uses of location data, showcasing what can be done, and pushing the industry forward. These partnerships have also been shown to generate positive public relations and marketing for the company.

Dr. Zhao indicated that Gravy Analytics data allowed her to answer questions using novel methods. Previous methods in disaster evacuation studies involved sending surveys to the people impacted by a disaster after the fact. This sampling method created many problems with accuracy and sampling bias. Using GPS data to map population movement during a natural disaster is more accurate and offers new insights for how to increase public safety and possibly reduce future injuries or fatalities caused by natural disasters.

Dr. Matz commented that attempting to collect or generate the data she received from Gravy Analytics on her own would have been nearly impossible and taken much longer. Dr. Matz flagged a benefit of using data controlled by companies: very few people will have conducted research on it. Publishing in peer-reviewed journals, a necessity for many academics, generally requires researchers to explore novel ideas or methods. In disciplines forced to rely on a limited set of open data, it's challenging to find an idea or method that someone else hasn't already published about. Until there is more open data for researchers to work with, corporate data sharing offers the best chance for those researchers to investigate new data, publish their results, and advance their discipline.[9]

Dr. Dahbura noted the importance of accurate epidemiological modeling, not just for initial public safety decisions in a pandemic but also because research suggests that the public quickly loses trust in epidemiology after a single inaccurate forecast. Once epidemiology loses public trust, people are less likely to adhere to public health precautions in the future. Dr. Dahbura maintains that his team's research is a great example of how sensitive information like location data is essential for public safety, notwithstanding the potential risks of using the same kind of data in other ways. Dr. Dahbura sees his work as attempting to develop precision public health efforts, a cousin of precision medicine, to produce community-specific mitigation protocols.

PARTNERSHIP INFORMATION

Gravy Analytics: gravyanalytics.com

Gravy Analytics, Data for Social Good Program: gravyanalytics.com/data-social-good

Dr. Xilei Zhao, University of Florida: www.essie.ufl.edu/people/name/xilei-zhao

Dr. Sandra Matz, Columbia University: sandramatz.com

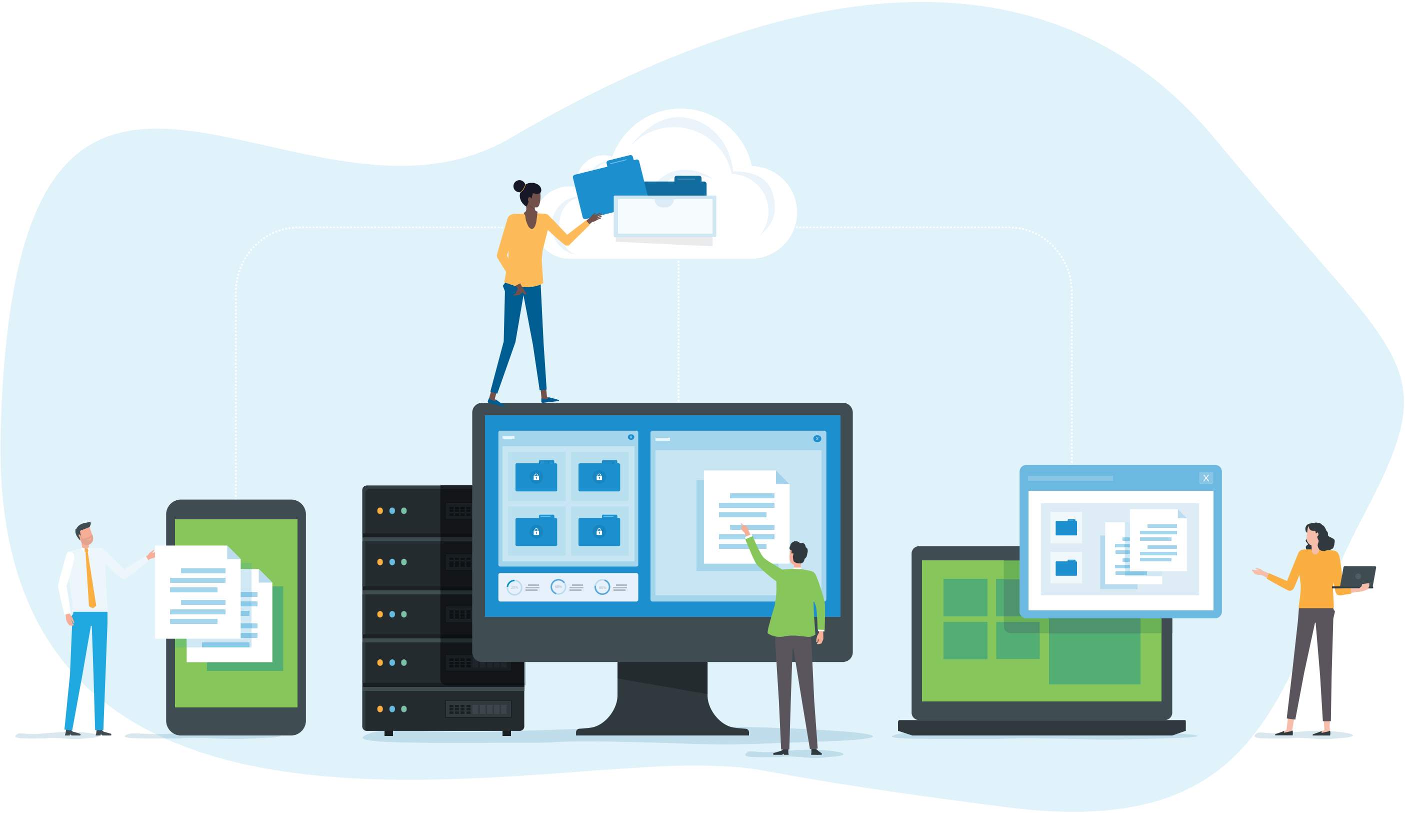
Data Science Institute at Columbia University: datascience.columbia.edu

Johns Hopkins University: www.jhu.edu

Dr. Anton Dahbura, Johns Hopkins University: engineering.jhu.edu/faculty/anton-dahbura/

Research papers from the partnerships:

- Estimating wildfire evacuation decision and departure timing using large-scale GPS data
- A highway vehicle routing dataset during the 2019 Kincade Fire evacuation
- Wildfire evacuation decision modeling using GPS data



IBM

Data Sharing Type
Internal, Closed Trusted Partnerships, Open Data

Organization and Partners

Company

With over 288,000 employees, International Business Machines Corporation (IBM) is a multinational technology company that provides a variety of computing and communications technologies and services for businesses. IBM is the largest industrial research organization in the world, with 19 research facilities globally. In 2022, IBM reported an annual revenue of $60.5 billion.[10]

Researchers

IBM hosts an in-house research organization composed of data scientists and researchers who process data to train models and improve services, among other projects. Externally, and in accordance with applicable data and privacy laws and IBM policies and practices, IBM researchers may share data for the same or similar purposes with universities, non-profits, and research labs around the world.

IBM researchers generally limit data sharing with third parties to non-sensitive data and metadata for research purposes and as part of data sharing partnerships. For example, working with the Australian utility Melbourne Water, IBM collected and processed data to develop insights that will help cut energy emissions. In limited instances when IBM shares data it has collected that includes personal information, IBM uses Privacy Enhancing Technologies (PETs). During the beginning of the global COVID-19 crisis, IBM collaborated with researchers and scientists to process SARS-CoV-2 genomic sequences, resulting in more than three million sequences, which were made available in a repository for researchers working to identify molecular targets for drug design, test development, and treatment. IBM is also working on big data machine learning projects using de-identified medical data (i.e., with personal identifiers removed) to advance scientific discoveries on disease progression, including diabetic kidney disease.

Partnership Considerations

Data

Representatives stated that, in limited scenarios, IBM may share a variety of non-sensitive data and metadata externally, depending on the purpose and nature of the research request. Data is shared only for the original purpose for which it was acquired, which can be found in each data acquisition's procurement statement. If IBM seeks to share data that includes personal information, before sharing, they use PETs to remove personal identifiers or render the datasets into a form that no longer constitutes personal data. IBM prioritizes PETs such as federated learning and differential privacy and has made libraries and toolkits publicly available.

Data Sharing

IBM faces ongoing demand for data sharing from inside the company, notably people in IBM Research, which supports a network of international research facilities and about 3,000 researchers. According to IBM representatives, when a researcher seeks access to third-party data, the researcher develops a proposal that is analyzed by procurement or contract professionals and counsel, who might request modifications to the governing terms and would be involved in any negotiations. Counsel, with support from IBM Privacy Office professionals and automated internal processes, analyzes the proposed uses to determine that they are consistent with the purposes for which the data was acquired, ensures compliance with IBM requirements and privacy implications, and addresses other sensitivities associated with requested data. That might include a check on the appropriateness of or permissions associated with data collected by IBM. These processes are designed to address issues of data privacy, security, and quality.

Company representatives reported that IBM researchers sometimes provide data externally through contributions to data sharing communities and sometimes directly to third-party partners in connection with an initiative, with a preference for open terms, in situations where non-sensitive and quality data and metadata is being shared that do not favor particular users or uses. For example, IBM worked with UK Research and Innovation — the UK government agency that directs research and innovation funding — to make available under open terms certain wave-elevation data. IBM favors the Community Data License Agreement (CDLA) permissive licenses for sharing open data. Unlike other open-sharing mechanisms, the CDLA permissive license is adapted to data sharing. For open data, IBM employs a review process to ensure that no sensitive data, such as personally identifiable or health-related data, is shared and that none of the data would be subject to privacy regulations. All IBM data sharing is based on a jurisdictional approach that accounts for differences in location and legal regimes — as a multinational company, IBM is accustomed to tuning its compliance based on jurisdictional requirements.

Privacy and Ethics

Representatives stated that IBM prioritizes transparency, data stewardship, privacy, security, and ethics, implemented through a technology ethics and privacy-by-design review process. Its AI Ethics Board enables IBM to take a centralized and multi-disciplinary approach to technology ethics. Their review considers the entirety of the use case, including the uses of data, such as training AI models where the risk of bias is a known concern, and it identifies methods for mitigating harm. In the limited instances where IBM researchers share personal data with third parties, the quantities are small, and agreements specify purposes, required consents, protections for privacy and security, and other terms. Throughout the formal review process, data sharing arrangements are subject to data constraints. A primary data constraint is the use of PETs (such as masking, encryption, or anonymization tools), which depend on the data's type, size, intended use, and source. Additionally, all arrangements are subject to IBM's security provisions.

Costs

IBM has both ongoing and fixed costs in privacy and security related to data sharing. These include staff time and internal tool development.

Next Steps

Representatives noted that because IBM is dedicated to continuing its data sharing arrangements, it is investing in related data protection, privacy, and security. IBM has developed a tool to systematize the review of third-party data to identify, among other attributes, personal information, and is specifically dedicated to increasing data sharing in the environmental, sustainability, and artificial intelligence contexts. IBM representatives stated they want to see more data sharing, particularly in the open data space. Former IBM executives were also involved in establishing an industry group, the Data and Trust Alliance, dedicated to improving data stewardship and potentially creating a vehicle for fostering practices supporting data sharing for research.

Risks and Benefits

According to IBM representatives, data misuse can cause significant harm. IBM continuously monitors ongoing lawsuits related to data and data sharing and reassesses risks as the landscape changes. Despite those risks, the benefits of data sharing to the company and the general public justify the practice, which IBM is proud of. IBM credits data sharing as having improved many company products and services. Executives believe corporations, governments, and citizens will profit from data sharing and open data. In the AI sector, for example, they contend that increased data sharing could help make AI models more equitable.

PARTNERSHIP INFORMATION

IBM: **www.ibm.com**

IBM Research: **research.ibm.com**

IBM Developer Datasets: **developer.ibm.com/exchanges/data/all**

Data and Trust Alliance: **dataandtrustalliance.org**

JOHNSON & JOHNSON

Data Sharing Type
Internal, Intermediated Data Sharing

Organization and Partners

Company

Johnson & Johnson is a multinational company specializing in pharmaceuticals, medical technology, and consumer healthcare with more than 152,000 employees and reported adjusted net earnings of $27 billion in 2022.[11]

Data Intermediary

The Yale University Open Data Access (YODA) Project is a data intermediary that facilitates clinical research data sharing. The YODA Project is located in the Center for Outcomes Research and Evaluation at the Yale School of Medicine.

Partnership Considerations

Calls for Transparency

In 2013, the European Federation of Pharmaceutical Industries and Associations (EFPIA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) jointly issued Principles for Responsible Clinical Trial Data Sharing, a report that aimed to spur more researcher access to information about clinical trials. The report was written in response to calls for greater transparency from pharmaceutical companies to ensure that drugs are safe and effective for the public. Since then, qualified scientific and medical researchers can request patient-level data for medicines approved in the U.S. and EU. As part of their effort to comply with these principles, Johnson & Johnson sought an independent organization to facilitate clinical trial data sharing with external researchers.

Johnson & Johnson and the YODA Project

The Yale Open Data Access (YODA) Project was selected as the independent review board between Johnson & Johnson and external researchers seeking anonymized clinical trial data. While there are additional mechanisms through which Johnson & Johnson shares data, the YODA Project is used for independent access requests to clinical trial data, resulting in over one hundred research publications documenting how novel research questions were answered with the analysis of data held by the YODA Project.

According to company and YODA Project representatives, Johnson & Johnson makes anonymized clinical trial data available for sharing through the YODA Project 18 months after study completion (allowing study investigators to publish first). Researchers submit research proposals to the YODA Project in order to request permission to access the data from Johnson & Johnson clinical trials. The status of the YODA Project as a separate entity from Johnson & Johnson supports the scientific integrity of the research. Because the researchers only interact with YODA Project personnel and processes, Johnson & Johnson can't restrict sharing data with researchers for any reason or improperly influence the findings of the research. All data requests are blinded to both the YODA Project and Johnson & Johnson during the request review process so that all researchers and institutions are considered on the basis of the merit and clarity of the proposal. While Johnson & Johnson requires researchers to share a copy of their manuscript with the YODA Project upon submission to a peer-reviewed journal, the company does not have the right to weigh in on the substance of the manuscript and has no decision rights in publishing. The YODA Project has supported data sharing efforts made by other companies such as Medtronic, and its funding model involves companies covering the costs of the initiative's expenses.

The YODA Project Data Sharing

YODA Project representatives communicated that they developed several methods that safeguard patient privacy, increase the likelihood of ethical use, and increase transparency about what data researchers can request. First, the YODA Project has standard and detailed Data Use Agreements that are publicly posted for researchers to see at any time. Second, the YODA Project provides a policies and procedures document that describes the full scope of data sharing so that the data requestors know what to expect at every stage, including data availability, requirements, internal and external review processes, due diligence assessments, data use agreements, and data distribution. Third, a significant feature of the YODA Project is that clinical trial data sets that research sponsors have agreed to share are publicly listed, and the interface and study metadata support both searching and browsing functions. Fourth, if a researcher doesn't see a known clinical trial data set they had hoped to find, they can submit a request to the YODA Project to determine whether the data can be made available. Fifth, the YODA Project maintains a list of clinical trials they can't share with the public for reasons such as the trial being incomplete, regulatory approval being pending, or the trial being older and the data hasn't been digitized. Lastly, the YODA Project provides a dashboard with metrics about their data sharing. The YODA Project staff works with researchers to clarify and strengthen proposals where needed.

YODA Project representatives added that data provided to the YODA Project for sharing are not transferred directly to researchers after a request is approved. Instead, they are made available through a secure analysis environment and accessed through a VPN. Company representatives said that Johnson & Johnson incurs a considerable expense to support the associated infrastructure. After accessing and analyzing the data, researchers can download their analyses but not the data themselves, and researchers must agree to several privacy-protective measures to avoid re-identifying patients who participated in the clinical trials. An illustration of the kind of research this partnership produces can be found in the systematic review prepared for the World Health Organization by Dr. Lawrence Mbuagbaw, an associate professor at McMaster University. Dr. Mbuagbaw used data held in part by the YODA Project to review the evidence and efficacy of bedaquiline for treating multidrug-resistant tuberculosis, ultimately informing global policy guidance on the treatment.

Clinical Data Preparation

Johnson & Johnson reported that it has an internal group that prepares data, performs assessments of quantitative risk for anonymization, and leads in the development of anonymization techniques at the company. Once a Data Use Agreement is fully executed between the YODA Project and the approved researcher's institution, researchers access data through an independent secure platform, where they can work on the data with embedded analytical tools, and then export the analysis. They also have the option to securely upload their own data to the platform in order to combine data sources.

Johnson & Johnson says their approach is to try and increase the utility of the clinical data as much as possible without compromising patient privacy. However, they recognized the need for a framework to identify potential risks and effective mitigation strategies that don't compromise the data's utility once its sensitivity reaches a certain threshold. Johnson & Johnson's Head of Clinical Data Standards & Transparency, Stephen Bamford, co-authored a paper titled 'Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data' where he and his co-authors explored a process to grade data depending on its utility on a scale from 0-5 depending on the amount of anonymization needed to suit a research method.

When preparing clinical data for sharing, Johnson & Johnson stated they use a variety of privacy techniques, such as minimization, key-coding, pseudonymization, anonymization, and clinical data synthesis, which creates a synthetic model to generate artificial but realistic study data. These different techniques allow Johnson & Johnson to produce different versions of data for different studies depending on the requirements. While pseudonymized or key-coded data is never let outside Johnson & Johnson's secure environment, there are some circumstances where anonymized or pseudonymized data can be shared, including through approved YODA Project researcher proposals.

Risks and Benefits

Risks

Johnson & Johnson representatives acknowledge that there are risks associated with data sharing depending on how that sharing is stewarded, including risks to individuals' privacy, to the obligations the company made in the informed consent process, and to the reputation of the company. However, the company believes that the benefits of data sharing significantly outweigh the risks, noting that none of the worst-case scenarios that were predicted in early discussions of data sharing have come to pass, in part because of the policies and processes that have been put in place by Johnson & Johnson.

Benefits

Company representatives believe that sharing existing data has enabled researchers to answer many novel research questions without exposing patients to the inherent risks of clinical trials. Johnson & Johnson also engages in less sensitive data sharing for research when appropriate. For example, the company recently contributed data to an antimicrobial resistance surveillance data register hosted on the Vivli platform. The data shared were from past clinical trials where participants from diverse geographical regions submitted sputum to be tested against pathogens. The project required minor data minimization due to the nature of how data were collected during the study and are hoped to yield meaningful public benefit.

PARTNERSHIP INFORMATION

Johnson & Johnson: **www.jnj.com**

The YODA Project: **yoda.yale.edu**

KHAN ACADEMY

Data Sharing Type
Internal, Closed Trusted Partnerships

Organization and Partners

Company

Khan Academy is a U.S.-based nonprofit organization founded in 2008 that operates a website and related applications providing online educational programming for students through instructional videos, online exercises, and instructional articles. Khan Academy has approximately 230 employees and reported revenues of over $59 million for the fiscal year 2021.[12]

Khan Academy and Formative Assessment Partners

Khan Academy representatives said they partnered with a third-party formative assessment provider to co-develop an educational tool based on student assessment scores. As part of their product-development partnership, they also developed a research partnership to understand how students' use of Khan Academy products affects student assessment scores across different demographics. The research project was co-developed by Khan Academy and its assessment partner and used a secure data warehouse to facilitate researcher access to shared, de-identified data.

Khan Academy and Standardized Test Partners

Another data sharing model has been used, at a somewhat smaller scale, with standardized test provider partners. Khan Academy offers test preparation courses for different standardized tests. Providers of such tests wanted to understand the relationship between student use of Khan Academy test-prep materials and student scores on standardized tests. The sample for these studies is smaller than the potential population using the test-prep courses. This is because the testing partner only requested consent from test takers and not from all users; additionally, not all test takers provided consent. Once the testing body sent Khan Academy the user list and confirmation of consent, Khan Academy queried the relevant data and securely shared it with the testing body, which merged in test scores and de-identified the record to complete their analysis. For the publishing phase, Khan Academy staff reviewed the researcher's drafts and provided feedback.

Khan Academy and School District Partners

In 2017, Khan Academy started partnering with school districts to measure how students' use of Khan Academy affects their scores on a standardized state test. In this data sharing model, the school district provides consent and direction for Khan Academy to share identified student usage data with the district securely. The district merges in student test scores and a subset of demographic data, de-identifying before securely sharing the full dataset with Khan Academy. Khan Academy then stores the data in a secure data warehouse, conducts the analyses, and shares the findings with the district. Khan Academy has worked with several school systems on research involving test scores from their districts and prefers long-term partnerships instead of one-time data requests. An example of published research from this type of partnership can be found in a company report titled Use of Khan Academy and Mathematics Achievement.

Partnership Considerations

Company Data Sharing Team

Khan Academy's Efficacy & Research team comprises three full-time people supporting data sharing partnerships as part of their overall responsibilities. This particular team's larger scope focuses on research into the efficacy of instructional techniques and often collaborates with external organizations or researchers. The team aims to develop queries that can be reused across different partnerships' data requests, thus reducing repetitive work. They have created clear data dictionaries so that partners can accurately understand shared data, and they offer some consultative support for data sharing partners.

Data Sharing

Khan Academy collects data for internal and external research and analysis through the operation and use of its platform. As part of its educational mission, Khan Academy is particularly interested in research to understand how the use of Khan Academy's learning platform affects mastery of the subject matter and student outcomes. In connection with providing its services, Khan Academy seeks out opportunities to partner with school districts and others to advance its research program. Districts that participate in research studies involving the use of test scores or other assessment data provide consent for assessment data to be shared with Khan Academy for efficacy research. Its research efforts generally focus on studies conducted in conjunction with its school district customers and other trusted partners. Khan Academy occasionally supports external research conducted by universities but generally declines third-party researcher requests, given the labor-intensive process required to curate fit-for-purpose data and negotiate data sharing agreements (DSAs). Successful partnerships involve identifying a dedicated counterpart at the partner organization with whom to negotiate the DSA and associated expectations, requirements, and terms.

Risks and Benefits

Khan Academy's legal framework addresses data governance, including privacy and security. The company expects any external research using its data to be under the auspices of an ethical framework, such as an Institutional Review Board (IRB) authorization, and conducted using de-identified data sets. External research partners enter into DSAs with terms that vary depending on the use of data and type of study. DSAs typically address the research goals, roles of each party, secure processes for preparing, transmitting, de-identifying, and storing the data, limits on the use of the data, and expectations regarding the publication or sharing of findings. In order to protect student data privacy, the company typically shares only de-identified data with research partners. An exception to this is data sharing with school districts. At the request and direction of the school district, Khan Academy will share the school district's own identified data, which the school district typically uses to merge with assessment data prior to de-identifying the research record. Moreover, Khan Academy applies data minimization principles and shares only a subset of all collected data. These partnerships publish research based on the shared data either jointly or independently. Regardless of who is publishing, disaggregated data is never shared in a publication.

PARTNERSHIP INFORMATION

Khan Academy: **www.khanacademy.org**

LINKEDIN

Data Sharing Type
Internal; Closed Trusted Partnerships

Organization and Partners

Company

Founded in 2003, LinkedIn is an employment-centered social media platform with over 900 million registered users, reported an annual revenue of $13.8 billion in 2022, and has more than 21,000 full-time employees. In 2016, LinkedIn became a subsidiary of Microsoft.

Partnership Considerations

Data Sharing Partnerships

LinkedIn has a specialized team of data scientists and public policy managers who administer its Data for Impact program, which is the primary mechanism LinkedIn uses to share aggregated, anonymized datasets with external partners at no cost. According to company representatives, there are generally three forms of data sharing partnerships: project-based data sharing (often in the context of long-term institutional relationships), observatory sharing, and collaborative research. The first category is typically one-time data requests, and their interactions with the project teams are relatively quick, even if their relationships with the institutions are long. The second category is when LinkedIn provides regular delivery of new data to an external partner, with the value being the consistency and recency of data across time. The third category, by contrast, requires many in-depth consultations because they engage with institutional staff on building a new indicator or co-authoring a report. The research questions, goals, internal costs, technical requirements, and privacy protections influence whether LinkedIn agrees to share data and, if so, what that data sharing looks like. A typical partnership launch involves a meeting to understand the researcher's request and how those align with data and privacy considerations. Often, partners want more granular data than LinkedIn is willing to give.

Next, LinkedIn conducts an internal assessment of the privacy risks and methods needed to execute the request. If approved, LinkedIn delivers the data to the external partner and then works with them to ensure the requestor's intended analysis aligns with the shared data to ensure methodological quality.

LinkedIn representatives stated they receive project-based data requests through the Development Data Partnership and the Industry Data for Society Partnership. These are usually used when researchers intend to produce a research product such as a report, organizational strategy document, or peer-reviewed publication. Most of the communication between partners and LinkedIn is concentrated between the initial request and the data shipment and then at key points along the process toward publication. Data observatories and embedded data products operate differently. An example of a data observatory is LinkedIn's partnership with the Inter-American Development Bank's Labor Market Observatories or the German Federal Statistical Office (Destatis). LinkedIn has explored long-term research relationships with entities like the Institute for Employment Research (IAB). These partnerships require more coordination and investment but also provide potentially higher-impact research outcomes.

Data License Agreements

LinkedIn has a standard data license agreement (DLA) and will only modify it slightly, if necessary, based on the institution they are partnering with. The company's representatives commented that it would be difficult to develop unique DLAs for each institution, thus their use of a standard DLA as much as possible. Their DLA focuses on guaranteeing LinkedIn member privacy, meaning no personally identifiable information or data is ever shared with external partners, and it also requires LinkedIn to review research partners' drafts prior to publication to ensure data is being interpreted and used correctly.

Data Menu

LinkedIn offers a public-facing Data Menu that displays a list of datasets offered to external partners. Company representatives emphasized that datasets listed on the menu undergo several layers of review for quality and are continually supported. The current categories of data include 1. LinkedIn hiring rate (their most popular), 2. Career Transitions, 3. Skills Genome, 4. Skills Similarity, and 5. Skills Penetration. All data in the menu is aggregated, anonymized, and can speak to labor market dynamics in 80 countries. LinkedIn offers various scheduled refreshes on data depending on the indicator, and data availability changes over time. Recently added datasets include indicators for gender and the green economy. To support a service like the Data Menu, LinkedIn has a continual updating, aggregation, and review process for quality and privacy. Representatives claimed that the data menu integration process gives partners more confidence in the data, as these data sets have gone through additional internal vetting and have been used by other partners. They also communicated that they would like to move towards more automation for data sharing and are exploring using differential privacy and synthetic data to help assure that no individual's data can be re-identified.

Data Sharing Capacity

LinkedIn representatives shared that they might be interested in data sharing more frequently, but only if doing so was tied to positive social or economic impact, and the company could maintain user privacy. They said they want the team to be the right size for the requests they receive. Getting additional investment in data sharing would also require justification. Several factors decide if a data sharing project is justified, including formal criteria such as the project's feasibility, potential impact, additive effect (asking if existing data could accomplish the same thing), and thematic relevance (asking if data sharing contributes to equity, sustainability, or resiliency). Several things can inform how impact is measured, such as the number of downloads, views, or citations a research product using LinkedIn data receives, better-informed decisions regarding global economic development policy, or influencing the future of employer training and skill development. LinkedIn is also exploring using an API for data sharing so researchers and policymakers can pull aggregated indicators without needing direct staff support.

Data Sharing and Privacy

There are several people in the company that help calibrate the right level of privacy safeguards and data granularity for sharing, including data scientists, economists, and policy experts. The data sharing team said they err on the side of caution, observing that there are generally two reasons they choose not to share data. First, if there are any privacy concerns within a data request that cannot be mitigated, and second, if the external partner is ill-equipped to understand the statistical limitations of the available data. Sometimes LinkedIn data can be complex, incomplete, or unsuited for the statistical methods the partner wants to use, and in those cases, LinkedIn does not share data with them.

Costs

Data storage, computation, IT infrastructure, and legal support were all listed as ongoing costs for data sharing, but the biggest cost for the company is the staff time to manage the program. LinkedIn reiterated that the cost of staff time is why they have strong DLA policies that reduce the negotiation period when onboarding new partners. The high cost of bespoke data sharing requests motivates their focus on developing long-term solutions and automated data sharing techniques.

Risks and Benefits

Risks

LinkedIn identified legal, reputational, intellectual property, and privacy violations as potential risk areas when sharing data. The company minimizes risks by only sharing aggregated, anonymized datasets with trusted public benefit partners who have signed DLAs. It manages remaining risks as effectively as possible by being transparent with members about the risks and mitigation techniques when members consent to share their data. They noted that partners occasionally mischaracterized or misinterpreted LinkedIn data in draft research outputs. When this happens, LinkedIn has to go back and meet with the researchers to rectify the error before publication. Lastly, there can be perception risks related to data sharing for public benefit. It can be challenging to effectively convey the benefits of data sharing with the public when there is reasonable public mistrust of data-collecting institutions.

Benefits

LinkedIn representatives said that the company's vision is to create economic opportunity for every professional in the world and that data sharing with external researchers and policy partners helps LinkedIn achieve that vision. Additionally, data sharing partnerships complement the data analysis done by its internal researchers. The representatives also conveyed that data sharing has led to unexpected ideas, creativity, and learning opportunities. For example, the IMF provided feedback about LinkedIn's skills data that gave the company insight into using their data in new ways.

PARTNERSHIP INFORMATION

LinkedIn: www.linkedin.com

LinkedIn Data for Impact: economicgraph.linkedin.com/data-for-impact

LinkedIn Data Menu: economicgraph.linkedin.com/data-for-impact#data

German Statistical Authority:
www.dashboard-deutschland.de/indicator/tile_1673880739519?mtm_campaign=dd-social-sharing

Inter-American Development Bank Labor Observatory: www.iadb.org/en/news/inter-american-development-bank-and-linkedin-join-forces-jobs-recovery-region

Organisation for Economic Co-operation and Development Artificial Intelligence Policy Observatory:
oecd.ai/en/data?selectedArea=ai-jobs-and-skill

META

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 third 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 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

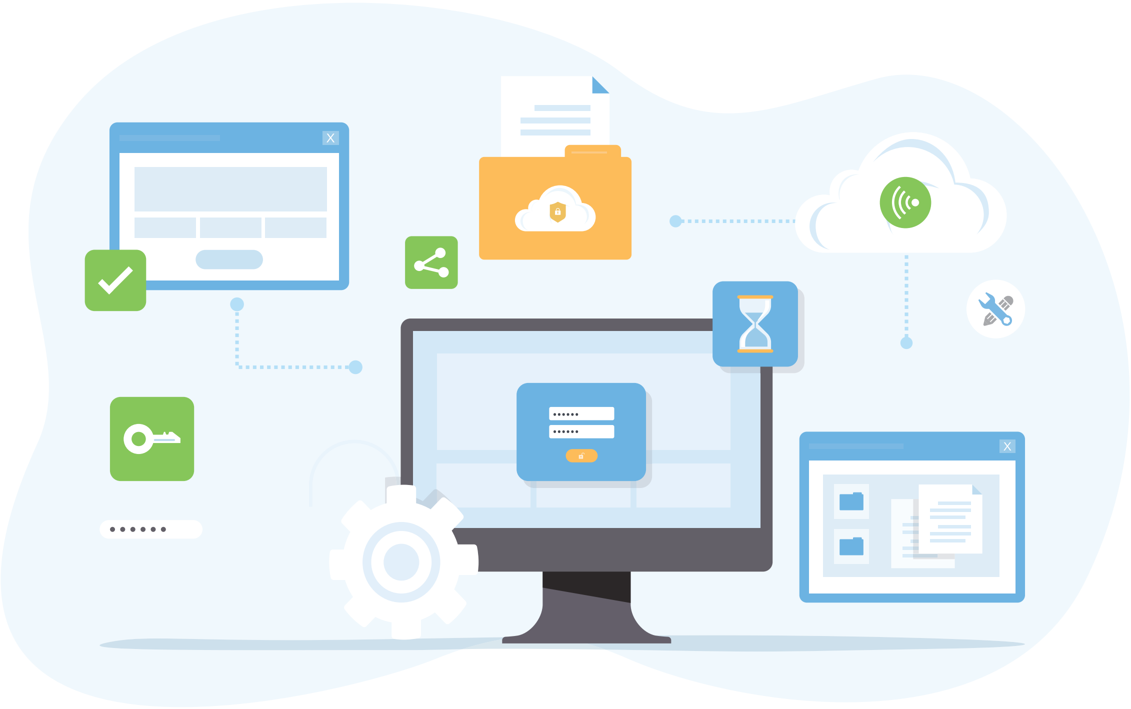
Meta: **about.meta.com**

Meta — Illustrative list of publications from data sharing partnerships: **developers.facebook.com/docs/url-shares-dataset/featured-works**

Meta — Data for Good: **dataforgood.facebook.com**

Meta — CrowdTangle Data for Researchers: **help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers**

U.S. 2020 Election Project: **research.facebook.com/2020-election-research/**



MICROSOFT

Data Sharing Type
Closed Trusted Partnerships, Open Data

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.[13]

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 U.S. 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 re-identification 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 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."[14] 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 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.[15]

PARTNERSHIP INFORMATION

Microsoft: **www.microsoft.com**

Industry Data for Society Partnership: **www.industrydataforsociety.com**

Answer ALS: **www.answerals.org**

United Nations International Organization for Migration: **www.iom.int**

CASE STUDY THEMES AND ANALYSIS

Although each partnership covered in this report had distinctive approaches to and experiences with sharing data for research, several themes emerged across the interviews.

Data shared by companies varied by type, influencing research use.

The companies and organizations interviewed, which differed by industry sector, size, and mission, each collected or generated a variety of data that was considered useful for research. The majority of data that was shared with researchers was created through the course of operating a business, what is sometimes called "administrative data" in academic literature.[16] Whether companies chose to make specific data available to researchers and under what conditions depended on the degree of its sensitivity, which could relate to privacy implications for individuals or its proprietary value. Access to data provided new opportunities for researchers to analyze new questions or revisit and update existing research. Disciplines with a scarcity of open data often have an oversaturation of research on existing datasets — people work with what they can get. For example, the Enron Corpus is a set of 600,000 emails resulting from the Enron Corporation's collapse and subsequent investigation of their email server. The corpus represents one of the few public collections of mass emails that researchers can study, which has led to an inordinate amount of published research using the corpus. Companies with textual data could make particularly impactful contributions to research if they shared text-based data.[17]

Companies and researchers take steps to protect privacy.

Most of the data shared in our case studies was privacy-sensitive, particularly in the biomedical, education, and location sectors. Both companies and researchers communicated that they use several techniques to protect the privacy of data subjects. The techniques companies reported using before sharing data with researchers include de-identification, aggregation, minimization, pseudonymization, K-anonymization, differential privacy, key-coding, confidential computing, restricted access permissions, VPNs, and secure environments such as data clean rooms.[18] All companies interviewed mentioned the use of cybersecurity techniques (e.g., encryption or trusted execution environments) in support of privacy protection. IBM specifically emphasized the privacy implications of metadata, even analyzing metadata separately in its internal privacy reviews. Though many steps were taken to protect data privacy before it was shared, researchers communicated that they often needed to take additional steps to protect privacy before their results were published, often using a subset of the techniques that companies used.

There is a need for ethics and privacy review for industry research.

Most of the partnerships included a step where a university researcher submitted an inquiry to their Institutional Review Board (IRB) for guidance on ethical research, privacy practices, and risk management. IRB approval is the standard benchmark for ethical research practices in academia. However, most IRBs are only available to university students and employees. All the companies in the case studies had to develop internal review processes to address potential issues, including research ethics, data privacy, risk assessment, legal compliance, user consent, and other decisions before sharing data with researchers. Every internal company review process that FPF analyzed differed in key respects, such as the party responsible for reviews (legal personnel, data scientists, policy, etc.), how long a review took, and evaluation criteria to determine appropriate privacy protections.

Additionally, while IRBs are a helpful support system for ethical research in academia, they are not adequate to address the complexity of corporate-academic data sharing partnerships. There is a clear demand for standardized, third-party, ethical and privacy review infrastructure for corporate-academic data sharing partnerships and industry research.[19] Key governance professionals, including Chief Privacy Officers, data protection staff, legal counsel, policy staff, and technical engineers are essential for enabling internal ethical and privacy-oriented review for corporate research until a better solution is implemented.

There are multiple ways to share data for research.

There were generally three types of data sharing modalities represented in the case studies: open data, such as in some Microsoft projects; closed trusted partnerships, such as those established by Gravy Analytics; and data intermediaries, such as the YODA Project mediating access to Johnson & Johnson data. In Open Data sharing, anyone can access the data on a public website, and there are usually fewer restrictions on its use compared to other modalities. Open Data approaches are mostly intended for data that has no risk of re-identification and is typically not about people. Closed trusted partnerships are where a company and an external party negotiate an agreement covering data sharing, and include many privacy, security, sharing, and use restrictions on the data. Closed trusted partnerships represented the majority of the data sharing by the companies interviewed. Data intermediaries involve an organization providing a third party with custodial responsibility over data. Before data is transferred from the data source to the data intermediary, they formally agree on the conditions for data access, but the intermediary administers all data requests. Data intermediaries have also arisen to support the analysis of shared data.

Data Sharing Agreements are essential for successful partnerships.

With the exception of Open Data sharing, every company interviewed shared their data using a data sharing agreement (DSA) or similar legal tool that researchers or third parties had to agree to before the data was transferred from the company. These DSAs were generally considered confidential and very few companies were willing to share the text of their DSA. There was a spectrum of approaches regarding DSAs: on one end, some organizations offered adaptability and negotiation with potential data sharing partners, while others were specific about templated uniformity regardless of the potential data sharing partner.

Companies with an adaptable DSA seemed to benefit from being able to partner with a greater number and variety of researchers or organizations. However, they commented that drawbacks included a longer negotiation period to arrive at an acceptable DSA, a more expensive negotiation process, and sometimes the implementation of bespoke engineering processes to package the data and ensure privacy. A standardized DSA appeared to enable a quicker, less expensive negotiation process for onboarding partners and a known process for handling data and privacy concerns. Nevertheless, some external partners couldn't meet the terms of the DSA and therefore were excluded from a data sharing partnership that otherwise would have benefited all parties involved.

Data sharing requires significant and ongoing costs.

Every company interviewed reported several ongoing costs related to data sharing, the most common being staffing, legal support, computation, IT infrastructure, data storage, transfer, security, and maintenance. Among these, staff time was credited as the most expensive aspect of maintaining a data sharing operation. One company stated that operations support for data sharing is an often overlooked expense including activities such as researcher onboarding, live support, and troubleshooting. In effect, all the expected costs that come with operating normal technical products apply to data sharing, and the start-up costs for data sharing are non-trivial.

Every company interviewed had multiple people who dedicated significant portions of their job descriptions to supporting data sharing activities and programs. These teams were all cross-disciplinary and required expertise in data manipulation, analysis, privacy, and policy. Some firms involved their marketing team in promoting published research and data sharing efforts. Other firms brought in software engineering and data scientists to support data packaging and analysis. Several companies or organizations described part of their job as mentoring researchers or helping with data analysis. The high degree of effort to maintain quality data sharing was referenced specifically with closed-trusted relationships; open data was not reported to require as much time or personnel, presumably because the key labor investment came during the process of making certain data open.

Data sharing has inherent risks, but risks can be managed.

Every person interviewed emphasized risks inherent in data sharing, generally described as falling within two tiers. Tier one includes direct privacy and security risks, such as re-identification, compromising the informed consent of the users, leaks, risks to intellectual property, misuse of the data (intentionally or unintentionally), or compliance breaches (especially for companies that operate internationally). Tier two includes secondary risks, such as threats to corporate or academic reputation, loss of grant funding or employment, or legal action.

Together these risks motivated the extensive use of mitigation techniques and, for university-based research partners, the use of Institutional Review Boards and the involvement of university lawyers were seen as additional risk-mitigation mechanisms. Notwithstanding these risks, every participant interviewed asserted that they were outweighed by the benefits of data sharing. A common sentiment [as expressed by one interviewee] was that "none of the worst-case scenarios that were predicted in early discussions of data sharing have come to pass." Executives from one company suggested that the cultural conception of risk needs to be reframed away from compliance-oriented fear and toward a more positive, social benefit-oriented approach, meaning that corporate definitions of risk should include identifying potential missed opportunities that could benefit people if companies don't share data.

There are technical knowledge and infrastructure gaps between companies and researchers.

Many of the organizations that shared data with academic researchers described a technical knowledge or infrastructure gap between the organization and the researchers that created barriers to sharing. For example, one company said their process for sharing data involved the use of a tool that the research team didn't have the capacity or knowledge to use. Another company that was considering how best to protect user privacy in a dataset ultimately decided against using a preferred set of privacy enhancing technologies because of inexperience with them on the part of the researchers that they were partnering with. These anecdotes point to an incongruence between the technical and infrastructure capabilities of companies and at least some researchers regarding data sharing.

Data sharing benefits researchers, companies, and society.

Several benefits of data sharing were raised by companies and researchers. These included the potential to positively inform data-driven policy and contribute to many sectors of research. It is difficult to overstate the degree of support that those interviewed had for the benefits of data sharing. Many communicated their belief that data sharing enables researchers to answer novel questions that may benefit corporations, governments, and citizens alike. Several participants said that sharing data with researchers has a secondary benefit of aiding the company's reputation. One surprising finding from the interviews is how many companies said that data sharing ultimately benefited the company's core products or services, often unexpectedly. Many companies gained valuable insights into new uses for their data and reported that the data sharing process led to new ideas, increased creativity, and additional learning opportunities. There was an implied consensus among participants that data sharing can help solve some of the world's most pressing societal challenges and make important contributions to research and society.

CONCLUSION

The case studies provide support for claims on the importance of data sharing agreements (DSAs), the potential benefits and risks of data sharing, and the costs of running data sharing programs. However, there were also three novel findings resulting from the project:

1. data sharing programs may benefit companies' core services or products in unexpected ways,
2. company-created, public-facing data sharing menus are an effective method for facilitating data sharing partnerships, and
3. there is a potential skill and infrastructure gap between companies and researchers regarding Privacy Enhancing Technologies.

As the landscape of corporate-academic data sharing continues to develop, more research is needed into these three new findings to confirm and expound on their nature. Data held by companies continue to be a potentially significant resource for researchers who can use it to expand the scale and scope of their research questions. While data sharing has inherent risks, they are generally known and can be mitigated. The benefits of data sharing to companies, research, and society present a compelling argument that data sharing for research is transitioning from being considered an experimental business activity to an expected business competency for established firms. The COVID-19 pandemic accelerated data sharing around the world. For some companies, it was their first experience doing so. If corporate culture as a whole decides to normalize data sharing for research and implement rigorous privacy protections, corporate data sharing partnerships will no doubt regularly produce solutions to critical social problems and simultaneously benefit the companies that participate.

ADDITIONAL FPF DATA SHARING RESOURCES

- The Playbook: Data Sharing for Research 2-Page Summary and Recommendations
- The Playbook: Data Sharing for Research-Full Report
- The Value of Responsible Data Sharing for Research: Infographic
- Award for Research Data Stewardship presented by The Future of Privacy Forum
- Future of Privacy Forum's Ethics and Data in Research Working Group
- Contract Guidelines for Data Sharing Agreements Between Companies and Academic Researchers
- Best Practices for Sharing Data with Academic Researchers
- Beyond IRBs: Ethical Review Processes for Big Data Research

APPENDIX

Methodology

FPF researchers sought out a diverse mix of companies representing different industries and motivations for collecting and sharing data for research. Researchers initially used convenience sampling by contacting companies that were FPF members. Not only do FPF members represent a variety of industries, but their membership also signals an interest in protecting privacy and a willingness to consider an invitation to participate in a case study about data sharing (although not all that were invited agreed to do so). Additional non-FPF member companies were approached based on FPF's existing connections with company personnel.

Information was elicited from company representatives (executives and/or staff) using a list of questions (see appendix) sent to the participants in advance, followed by a semi-structured discussion addressing background information and specifics about a company's data sharing with researchers. The questions explored high-level data sharing activities, such as how companies identify data sharing partners, what kind of data is shared, under what circumstances, and with what considerations of risks and benefits. Most discussions lasted an hour and were recorded when possible. Recordings were strictly to supplement investigator note-taking and not for publication. Internal experts spoke on behalf of companies and, when feasible, offered introductions to researchers with whom they shared data. Researchers who agreed to be interviewed were sent a modified questionnaire ahead of time (see appendix) and were recorded for a semi-structured interview covering broad themes about their experience conducting research using data shared by a company.

In total, FPF contacted 34 companies, 13 of which were willing to participate in an interview. Four companies requested an informational, non-recorded meeting to ask questions about the nature and goals of the project before agreeing to participate. FPF interviewed 13 companies and eight research groups for a total of 35 people spanning 20 individual meetings. From those interviews, FPF determined there was sufficient information to form eight full case studies. FPF researchers used participant validation to increase the reliability of qualitative interview data. Interviewed participants were provided drafts of their case studies and were invited to correct, expand, or clarify text pertinent to their respective roles. Case studies went through two rounds of participant validations on average.

This process, like most qualitative research, yielded case studies that vary in scope and length. All case studies focus on the perspective and experiences of the participants, with the goals of illuminating business and research practices and encouraging more companies to share data with more researchers. Although FPF sought to interview company representatives, researchers from outside the company, and other stakeholders as appropriate, some case studies only feature company perspectives because it proved infeasible to contact or consult others. Moreover, given the intrinsic differences in the way companies share data for research, each case study was shaped to make the most of the information collected and to document the range of experiences. Finally, FPF learned from conversations with several companies for which it did not have enough information to develop a suitable case study (Uber, 3M, Plaid, Education Analytics, and Comcast), but the report's insights draw from the information they shared.

Interview Questions

For Companies

1. Please share your name and role.
2. Could you please describe the process of sharing data with researchers?
3. Do you have a standard data sharing agreement?
4. Are you able to share it with us?
5. How often do you share data with researchers (rarely, on occasion, a lot)?
6. Is the current data sharing you're doing the right amount, or would you wish to share data more or less often than you do?
7. How and how much do you constrain the data you share with researchers (not at all — public posting/API, selected data only but broadly available, selected data negotiated on a project-by-project basis)?
8. If data sharing involves sensitive data, what measures were taken to protect privacy or confidentiality?
9. By whom?
10. What costs have you experienced from sharing data with researchers (time of key personnel, IT, legal, internal research, data storage and communication, other)?
11. Were those costs limited to making arrangements for the first time or do they persist?
12. Do you think there are any risks to data sharing (legal, reputational)?
13. What benefits have you experienced from sharing data with researchers (Direct for your business, indirect via reputational boost, other)?
14. Do you maintain–and are you willing to share–a list of published research that draws from data you have shared with researchers?
15. Is there anything else you think we should know?

For Researchers

1. Please share your name, role, and institutional affiliation.
2. How much does your research depend on getting data collected or generated by companies?
3. Have you tended to focus on a partnership with one or a few companies or do you seek to obtain data from many companies?
4. What strategies, methods, or technologies worked well for getting data from a company?
   - What could have worked better?
5. If anything was done to protect private information in the data from the company, how much of the protection was done by the company before you got access to the data, and how much did you do?
6. What benefits have you experienced from getting data from companies?
7. What risks have you experienced from getting data from companies?
8. What role has your institution played in supporting your research or partnership with the company?
   - Did you interact with your institution's general council or Institutional Review Board for this partnership?

ENDNOTES

[1] American Economic Association. 2022. AEASTAT: Administrative Data

Access to Federal Government Administrative Data

www.aeaweb.org/about-aea/committees/economic-statistics/administrative-data

[2] Harris, L., Charma, C. 2017. Understanding Corporate Data Sharing Decisions: Practices, Challenges, And Opportunities for Sharing Corporate Data with Researchers. The Future of Privacy Forum. November.

fpf.org/wp-content/uploads/2017/11/FPF_Data_Sharing_Report_FINAL.pdf

[3] Jordan, S., Arledge, E., Stepanovich, A., Swauger, S., Blumenthal, M., Auh, R. 2022. The Playbook: Data Sharing for Research. The Future of Privacy Forum. fpf.org/wp-content/uploads/2022/12/FPF-Playbook-singles.pdf

[4] Goroff, D., Polonetsky, J., & Tene, O. (2018). Privacy protective research: Facilitating ethically responsible access to administrative data. The ANNALS of the American Academy of Political and Social Science, 675(1), 46-66. journals.sagepub.com/doi/pdf/10.1177/0002716217742605?casa_token=Os5QrpKPwuEAAAAA:vFLnaTIakQJq9QaXxDQOhwmOnr9uvkcvoyr8v_uq0_bcpqCtdq18yPooX1TphN9jfenG1w3LbRE

[5] www.aimscollaboratory.org/

[6] The Future of Privacy Forum is supported in part by the Bill and Melinda Gates Foundation.

[7] www.cbinsights.com/company/gravy-analytics/financials

[8] The Open Science Framework (OSF) is a platform for sharing the products of a research lifecyle such as data, protocols, study designs, reports, or publications. See osf.io/.

[9] Matz acknowledged the valuable support of the Data Science Institute (DSI) at Columbia. As outlined by DSI's Executive Director, Sharon Sputz, DSI has a lot of experience negotiating data sharing agreements and engaging the Columbia Sponsored Projects Office and lawyers. Sputz attends to the publication needs of researchers and their students, which often mitigate against non-disclosure agreements. She remarked on the tensions between research and privacy, a topic being addressed through an NSF grant.

[10] International Business Machines. '2022 IBM Annual Report.' 2022. www.ibm.com/annualreport/

[11] Johnson & Johnson. '2022 Johnson & Johnson Annual Report.' 2022. www.investor.jnj.com/asm/2022-annual-report

[12] Khan Academy. 2022. Khan Academy 2021-2022 Annual Report. s3.amazonaws.com/KA-share/annualreport/Khan_AnnualReport-22_R5.pdf

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

[14] 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.

blog.thegovlab.org/post/a-curation-of-tools-for-re-use

[15] 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. businesscase.opendatapolicylab.org/

[16] Goroff, D., Polonetsky, J., & Tene, O. (2018). Privacy protective research: Facilitating ethically responsible access to administrative data. The ANNALS of the American Academy of Political and Social Science, 675(1), 46-66. doi.org/10.1177/0002716217742605

[17] Administrative data that is not privacy-sensitive but could have proprietary value did not loom large among the companies studied, although it was a factor with others that were consulted: 3M shared specification and distribution data with government agencies and retailers in a joint effort to detect counterfeit N95 respirators; Comcast shared summaries of sampled network metadata supporting analysis of data flows for third-party network connections; and Uber has shared information about when it launched in different markets and such high-level statistics as the number of riders or drivers in a given market.

[18] See the 'Access Controls" section in The Playbook: Data Sharing for Research. fpf.org/wp-content/uploads/2022/12/FPF-Playbook-singles.pdf#page=30

[19] This finding confirms what FPF has previously recommended and is an indication that this challenge has remained unaddressed. See, e.g., Jordan, S., Arledge, E., Stepanovich, A., Swauger, S., Blumenthal, M., Auh, R. 2022. The Playbook: Data Sharing for Research. The Future of Privacy Forum. fpf.org/wp-content/uploads/2022/12/FPF-Playbook-singles.pdf

FUTURE OF PRIVACY FORUM

1350 EYE STREET NW | SUITE 350 | WASHINGTON, DC 20005   FPF.ORG | INFO@FPF.ORG