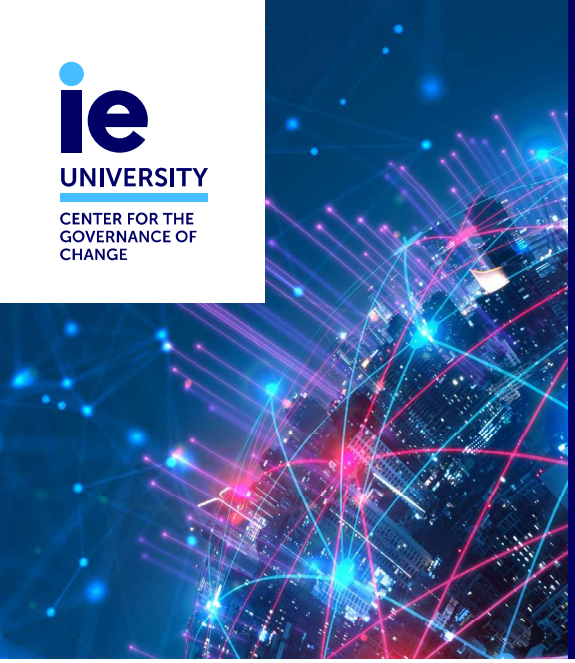

POLICY BRIEF: DATA COLLABORATIVES

Policy Brief
Center for the Governance of Change
September 2023

EXECUTIVE SUMMARY

Despite the abundance of data generated, it is becoming increasingly clear that its accessibility and advantages are not equitably or effectively distributed throughout society. Data asymmetries, driven in large part by deeply entrenched inequalities and lack of incentives by many public- and private-sector organizations to collaborate, are holding back the public good potential of data and hindering progress and innovation in key areas such as financial inclusion, health, and the future of work.

More (and better) collaboration is needed to address the many data asymmetries that exist across society, but early efforts at opening data have fallen short of achieving their intended aims. In the EU, the proposed Data Act is seeking to address these shortcomings and make more data available for public use by setting up new rules on data sharing.

However, critics say its current reading risks limiting the potential for delivering innovative solutions by failing to establish cross-sectoral data-sharing frameworks, leaving the issue of public data stewardship off the table, and avoiding the thorny question of business incentives.

This policy brief, based on **Stefaan Verhulst's** recent policy paper for the **Center for the Governance of Change**, argues that data collaboratives, an emerging model of collaboration in which participants from different sectors exchange data to solve public problems, offer a promising solution to address these data asymmetries and contribute to a healthy data economy that can benefit society as a whole. However, **data collaboratives require a systematic, sustainable, and responsible approach to be successful,** with a particular focus on:

Establishing a new science of questions, to help identify the most pressing public and private challenges that can be addressed with data sharing.

Fostering a new profession of data stewards, to promote a culture of responsible sharing within organizations and recognize opportunities for productive collaboration.

Clarifying incentives, to bring the private sector to the table and help operationalize data collaboration, ideally with some sort of market-led compensation model.

Establishing a social license for data reuse, to promote trust among stakeholders through public engagement, data stewardship, and an enabling regulatory framework.

Becoming more data-driven about data, to improve our understanding of collaboration, build sustainable initiatives, and achieve project accountability.

BACKGROUND: BIG DATA, DATAFICATION, AND DATA ASYMMETRIES

To understand the emerging problem of data asymmetries, it is useful to begin with the concepts of datafication and big data, a cross-sectoral phenomenon resulting from the widespread digitalization of our economies and societies.

Although it is typically understood as a quantitative issue, i.e. referring to the increasing abundance of information, big data extends beyond mere bigness. It is also about velocity, variety, and veracity, the three Vs of the Zettabyte era[i]. It is characterized by the proliferation of unstructured information, like images and audio, over standard formats that can be easily read by computers, such as text. And, crucially, it is defined by our ability to extract new meaning from the aggregation of such large disparate datasets.

Indeed, the rise of big data accelerated profound changes in the way information is collected, stored, and analyzed, which, combined with advances in computational capacity and an ever-greater reliance on data for decision-making, have contributed to the exponential datafication of almost every aspect of our public and private lives. Today, the proliferation of digital sensors and Internet of Things (IoT) devices, the virtually unlimited storage capacity enabled by Cloud and Edge Computing, and the advent of powerful artificial intelligence (AI) and data analysis tools, have resulted in a data ecosystem the scale and complexity of which boggles the mind.

Yet paradoxically, an era of plenty is also marked by scarcity, silos, and asymmetries. **The reality is that there is a critical mismatch between data supply and demand, as the information that could be most useful to decision-makers and relevant stakeholders rarely gets applied to the social, economic, and political problems it could help solve.** This stems from the growing hoarding of data by certain actors, which can be seen in the clear imbalances that exist between citizens and private- and public-sector organizations. Such asymmetries occur whenever there is a divide or disparity in the access to and re-use of data, and they can take up many forms, depending on the relationship between data holders, subjects, and users. Each of these manifestations poses unique problems, but considered together, they make clear the broader stakes of the challenge and suggest the need for stronger sharing within and among sectors.

DISCUSSION:
THE LIMITS OF PAST AND PRESENT DATA-SHARING EFFORTS

The pathway to more data sharing runs through the well-established (yet poorly understood) practice of open data. However, most early efforts at recognizing emerging data silos and easing the resulting asymmetries have thus far fallen short of their intended aims. For example, much of the released data has focused on national and supra-national organizations, even though larger amounts of data are held in silos at the subnational and local levels. Moreover, they haven't stopped large corporations, particularly in the tech sector, from hoarding users' data without transparently disclosing their intentions in order to gain an advantage over competitors and avoid attracting unwarranted attention from regulatory authorities. Lastly, these attempts have ultimately failed to enhance trust among users, who have grown increasingly concerned about the misuse of their personal information, and between data holders, who still largely avoid sharing data even with stakeholders outside their sector. The result is that many asymmetries have persisted, and today, much—perhaps a majority—of generated data remains locked away and inaccessible to those who need it most.

Figure 1: Data Act Problem Tree.[ii]

<table>
<tr><th></th><th>B2B/B2C DATA ACCESS</th><th>B2B DATA ACCESS</th><th>B2G DATA ACCESS</th><th>DATA PROCESSING SERVICES</th></tr>
<tr><td>PROBLEMS</td><td>Individuals and businesses often face challenges in fully harnessing the potential value of data produced through their utilization of goods and services.</td><td>In B2B interactions, there is a notable scarcity of accessible data, which hampers the ability to generate additional value.</td><td>Inefficient methods employed by the public sector when utilizing data from the private sector create a burdensome situation for companies.</td><td>Obstacles in transitioning between cloud and edge services, as well as concerns regarding unauthorized access to data by third countries.</td></tr>
<tr><td rowspan="3">DRIVERS</td><td colspan="2">Lack of legal clarity creates uncertainty for both consumers and businesses when it comes to accessing and utilizing data.</td><td rowspan="2">Insufficiently developed regulations and mechanisms for public sector entities when they need to utilize business data under exceptional circumstances</td><td>Unjust market practices and vendor lock-in in cloud and edge services</td></tr>
<tr><td colspan="2">Exploitation of uneven contractual terms in data access and absence of standardized data-sharing practices.</td><td rowspan="2">Possibility of data access conflicting with both EU and national legislation raises concerns regarding trustworthiness, security, and privacy.</td></tr>
<tr><td colspan="3">Lack of standards for reusing data within and among sectors</td></tr>
<tr><td>CONSEQUENCES</td><td>Limited competition, slower pace of innovation in data markets</td><td>Reduced options for consumers and increased prices for data-related products and services</td><td>Decline in EU competitiveness within the global data economy</td><td>Poor-quality public service delivery and a fragmented market</td><td>Inadequate computing resources for facilitating data sharing</td></tr>
</table>

> In Europe, the European Commission has sought to address these limitations with the recent Data Act, a proposal on harmonized rules on fair access to and (re)use of data that aims to make more information available for public use[iii].

The law seeks to establish new rights and obligations to ensure fairness in the digital environment, stimulate a competitive data market, and open opportunities for data-driven innovation, promising to unlock its untapped value across the EU. It has been hailed as the final and arguably most important element of Europe's ambitious digital transformation program, and it is currently undergoing interinstitutional negotiations (a.k.a. trilogues) with the Parliament and Council.

However, the proposal has received a fair degree of criticism, particularly from industry associations, due to its potentially negative impact on companies' data-driven business models and failure to increase legal certainty. The leading cause for concern relates to the protection of trade secrets, but the lack of clarity about the scope of the application and the limits on Business-to-Government (B2G) data sharing have also been highlighted. Another particularly sensitive issue is the absence of market-led compensation models that take into consideration the maintenance, technical, and administrative costs of making data available to third parties, as is the lack of adequate liability regimes and dispute settlement mechanisms to help protect consumers and increase trust.

Ultimately, the Data Act also risks falling short of its ambitious aims. The absence of cross-sectoral data-sharing frameworks to facilitate the combination of data from different sectors may stunt the potential for delivering new services for people and businesses. More public-sector data should also be made available to help organizations come up with innovative solutions to our most pressing challenges. Finally, incentives, monetary and otherwise, should be clarified to bring companies to the table and ensure they can enhance their capabilities over time. If we truly want to harness the potential of data to enable a healthy data economy, we need a new model of collaboration that can benefit society as a whole.

TOWARDS SOLUTIONS: DATA COLLABORATIVES

In recent years, one model has received increased attention from both public- and private-sector entities due to its potential to overcome the many bottlenecks and asymmetries that exist within today's data ecology and address our persistent failure to reuse data for the public good: data collaboratives. **The term refers to an emerging model of collaboration in which participants from different sectors exchange data to solve public problems,** by drawing together otherwise siloed data and a dispersed range of expertise, matching supply and demand, and ensuring that relevant institutions and individuals are able to use data in a way that maximizes innovative social solutions.

As they have moved from theory to practice, data collaboratives have spread around the world and brought large efficiency gains to various economic sectors. But although certain patterns are becoming clear, they are far from a uniform phenomenon. At least six different types of data collaboratives may be distinguished, each offering its own lessons and cautions for the ultimate goal of increasing (and improving) data sharing:

Figure 2: Typology of data collaboratives.

PUBLIC INTERFACES

Companies provide open access to certain data assets, enabling independent uses of the data by external parties.
Examples include Application Programming Interfaces (APIs) and Data Platforms.

TRUSTED INTERMEDIARY

Third-party actors support collaboration between private-sector data providers and data users from the public sector, civil society, or academia.
Examples include Data Brokerage and Third-Party Analytics Projects.

DATA POOLING

Companies and other data holders agree to create a unified presentation of datasets as a collection accessible by multiple parties.
Examples include Public Data Pools and Private Data Pools.

RESEARCH AND DEVELOPMENT (R&D) PARTNERSHIPS

Companies engage directly with public-sector partners and share certain proprietary data assets to generate new knowledge with public value.
Examples include Data Transfers and Data Fellowships.

PRIZES AND CHALLENGES

Companies make data available to participants who compete to develop apps, answer problem statements, test hypotheses and premises, or pioneer innovative data uses for the public interest and to provide business value.
Examples include Open Innovation Challenges and Selective Innovation Challenges.

INTELLIGENCE GENERATION

Companies internally develop data-driven analyses, tools, and other resources, and release those insights to the broader public.

However, it is also important to remember that data collaboratives—like any effort at data sharing—may also pose certain implementation challenges. Many of these obstacles were described in the preceding section as significant barriers to past and present data-sharing efforts, so they will need to be kept in mind when designing operational models that can truly unlock the untapped value of data. To ensure they can contribute to a healthy data economy that can benefit society as a whole, we must make data collaboratives systematic, sustainable, and responsible, with a particular focus on:

Challenge #1

Lack of Awareness and Data Literacy

There is a general lack of awareness about the opportunities for data reuse, as well as a more specific lack of understanding of the many instances where a particular dataset could be directed to solve a public problem or fulfil a consumer need.

Solution—A New Science of Questions

The sheer variety and complexity of challenges facing our world can sometimes be overwhelming. We know that (shared) data can help solve these challenges, but policymakers are often presented with problems of prioritization. Establishing a new science of questions can help identify the most promising public and private challenges and better understand what types of data should be shared, with whom, and through which mechanisms. To do that, policymakers should consider launching initiatives such as *The 100 Questions*[iv] to map the most pressing, high-impact questions that could be answered if relevant datasets were made available and develop new participatory methodologies to solve them by leveraging the power of experts who possess both domain-specific knowledge and data science specializations.

Challenge #2

Absence of Trust

There is a pervasive absence of trust both among potential sharing partners and the general public, who remain skeptical about how its data is being (re)used. While such concerns are understandable, the absence of trust acts as a barrier to the potential of data sharing.

Solution—Establish a Social License for Re-Use

As much as operationalizing collaboration depends on incentivizing data holders, its success ultimately rests on making the case more broadly to society at large. If data collaboration is to be operationalized at scale, then all stakeholders must be able to trust that all parties will uphold their responsibilities when it comes to how data is collected, stored, and used. To that end, policymakers ought to help establish a sort of "social license" for data collaboration and focus efforts on:

- Supporting public engagements, in the form of data literacy campaigns, citizen assemblies, or open dialogues with different stakeholders.
- Leveraging data stewards, in their role as conduits between various stakeholders and data practitioners.
- Establishing an enabling regulatory framework, to operationalize data sharing by building trust.

Challenge #3

Uncertainty within the Private Sector

Then, there is the issue of unclear incentives within the private sector, as concerns about data leaks, penalties, and reputational losses abound. Some of these concerns are no doubt legitimate, but they act as a barrier to unleashing the potential of data for the public good.

Solution—Clarify incentives

Sharing cannot rest on altruism alone, yet a lack of clarity surrounding incentives is one of the major impediments to greater data collaboration. In fact, concerns over perceived competitive threats or regulatory repercussions can actively disincentivize sharing, especially by the private sector. A market-led compensation model may address these concerns best and include provisions for reinvestments and innovation that ensure data-sharing organizations can enhance their capabilities over time. However, collaboration agreements should also highlight more intangible benefits, such as those developed by the author in the 9Rs *Framework*[v], including:

- Gaining access to data assets held by other organizations
- Rectifying errors and improving data quality
- Generating new answers to questions
- Enhancing an organization's image and reputation
- Attracting and retaining data science talent

Challenge #4

Limited Capacities

Additionally, many organizations have limited capacities in terms of their ability to process, analyze, and use data, and require significant investments to carry out their data collection and management tasks. These can inhibit their willingness to share and reuse data.

Solution—Profession of Data Stewards

A critical factor of effective data sharing is whether there exists within sharing organizations individuals or teams specifically empowered to initiate, facilitate, and coordinate data collaboratives. These are known as "data stewards," people who have the requisite expertise and authority to recognize opportunities for productive collaborations and respond to external data requests. More specifically, they perform six key functions:

- Partnership and community engagement
- Internal coordination and staff engagement
- Data audit, ethics, and assessment of value and risk
- Dissemination and communication of findings
- Nurture data collaboratives for sustainability
- Decarbonization of data assets

Challenge #5

Limited Community of Practice and Knowledge Base

Lastly, there is still a limited community of practice and a relatively weak knowledge base, which will need a more solid foundation to facilitate true collaboration among sectors. The nascent nature of data sharing poses an additional barrier.

Solution—Become Data-Driven About Data

For all the examples that suggest the power of data sharing, the evidence base remains thin, and both the theory and practice of collaboration are unsystematized and under-researched. This limits the reproducibility and scalability of data-sharing projects. Without more knowledge of what works (and doesn't), it is harder to establish best practices and operational guidelines that can help build sustainable and responsible data-sharing initiatives. And without a better understanding of impact, it is also harder to improve initiatives and achieve accountability for projects that cause direct or indirect harm. Thus, we need more data on things like:

- What are the kinds of projects that organizations seek data to address?
- What types of data are being shared, and how?
- What safeguards are being implemented to ensure adequate protections?
- What is the impact—positive and negative—of data sharing initiatives?

ENDNOTES

i The Zettabyte Era refers to a period beginning in late 2016 when the amount of Internet traffic exceeded one zettabyte (one trillion gigabytes) for the first time.

ii The information in the table is adapted/comes from the Impact assessment (SWD (2022) 34, SWD (2022) 35 (summary) accompanying a Commission proposal for a regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (data act) (COM (2022) 68).

iii The Data Act complements the 2022 Data Governance Act, another legislative measure aimed at boosting data sharing and European data spaces. Together, these make up the two main proposals released as part of the European Strategy for Data, which hopes to create a single market for data that will ensure global competitiveness and data sovereignty. They also support the EU's existing data framework, including the General Data Protection Regulation (GDPR), the Free Flow of Non-Personal Data Regulation, and the Open Data Directive, as well as the recent Digital Markets Act (DMA) and Digital Services Act (DSA).

iv In 2019, the GovLab, in collaboration with Schmidt Futures, launched The 100 Questions initiative. The initiative sought to establish priorities by mapping the world's 100 most pressing, high-impact questions that could be answered if relevant datasets were made available. Find out more here: https://the100questions.org/.

v Zahuranec, Andrew J. 2021. "The '9Rs Framework': Establishing the Business Case for Data Collaboration." Data Stewards Network (blog). November 9, 2021. https://medium.com/data-stewards-network/the-9rs-framework-establishing-the-business-case-for-data-collaboration-26585455ccc0#:~ .

ie UNIVERSITY

CENTER FOR THE GOVERNANCE OF CHANGE

This policy brief was produced within the framework of the Center for the Governance of Change's research program *The Digital Revolution and the New Social Contract*, and in particular its second work package, which studies the emergence and governance of the data economy, and how it can be fair, competitive, and safe.

It is based on the latest policy paper of the package, authored by Stefaan Verhulst, *"Data Collaboratives: Enabling a Healthy Data Economy through Partnerships"*, IE CGC, July 2023

You can access the paper and learn more about the program here: https://www.ie.edu/cgc/research/new-social-contract-digital-age/

Design: epoqstudio.com

cgc.ie.edu