COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

White Paper

Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data

July 2019

Contents

The views expressed are those of certain participants in the discussion, and do not necessarily reflect the views of all participants or of the World Economic Forum.

This white paper has been published by the World Economic Forum as a contribution to a project, insight area or interaction. The findings, interpretations and conclusions expressed herein are a result of a collaborative process facilitated and endorsed by the World Economic Forum, but whose results do not necessarily represent the views of the World Economic Forum, nor the entirety of its Members, Partners or other stakeholders.

Foreword

Genya Dana,
Head, Precision Medicine

Venessa Candeias,
Head, Shaping the Future of Health and Healthcare

At the World Economic Forum, we think of data as the oxygen that fuels the fire of the Fourth Industrial Revolution. It is readily available and necessary, but if used improperly, it can generate dangerous and unwelcome results. Concerns over how to protect valuable data, especially sensitive, personal data, are at the core of many countries' and institutions' data policies. We see a complex and dynamic data-policy landscape, where it is becoming more difficult to share data to the extent desired to advance research and innovation.

This paper aims to create a baseline understanding of federated data systems for policy-makers and leaders of the Fourth Industrial Revolution and to elucidate how such systems may help us navigate a variety of data-policy and governance challenges. Federated data systems are not new per se, but they are starting to be deployed more frequently as a solution to accessing multiplying, disparate data repositories in a multinational and multi-jurisdictional world.

Health-related data, one of the most sensitive types of data, is proliferating at a dizzying pace. Leaders in health and healthcare are in search of new models of data access that enable them to capture the value of such data and provide better patient care and drive innovation. The adoption of federated systems is one promising approach, offering a critical component in powering a sustainable, inclusive and transparent digital economy. We believe that the widespread use of federated data systems can stimulate innovation by allowing access to sensitive data while minimizing risks and unintended consequences. By raising appropriate questions that will help guide policy-makers and leaders, the hope is that this white paper raises important issues and ways to address them that must be considered when designing federated systems for accessing sensitive data in a scalable, trustworthy and legally compliant manner.

This paper is part of a Forum project in which the health and data-policy communities navigate the complex landscape of data policies via the use of federated data systems. Federated systems enable queries to be sent between disparate data repositories, or nodes in a federation. These travelling queries allow the data to remain localized, while providing access to data elements that are needed to accelerate research and innovation. The Breaking Barriers to Health Data project examines how to set up federations that enable access to disparate sets of sensitive health data via queries, starting with genomic data for rare disease research and diagnosis across four different countries.

Proactive efforts will be needed to motivate government officials, business leaders and civil society members to establish real-world pilots and to enable continuous and active experimentation with federated data systems, particularly in situations where they are most valuable. The technology is not the most difficult aspect of setting up such efforts; the governance of such systems and engendering a level of trust between nodes will be the more difficult challenge. Investing in an iterative process of multistakeholder dialogue, piloting, experience-sharing and refinement will be critical to ensure that federated approaches are human-centred, flexible and responsive to the complexities and dynamism of today's data-policy landscape.

Introduction

Networked digital technologies of the Fourth Industrial Revolution are quickly becoming the engine of change throughout all sectors of the global economy. By redefining the ways in which industries, individuals, institutions and governments all interact through data, there is a unique opportunity to create a more inclusive, innovative and resilient society.

The aim of this white paper is to contribute to the global discourse on data policy by advancing new approaches for balancing the need to protect sensitive data with the need to use it in innovative and trustworthy ways. Some of the most difficult data-policy questions relate to cross-jurisdictional concerns about protecting sensitive data while also ensuring its ability to offer new insights. This white paper aims to illustrate how federated data systems can strengthen trust and enable more effective data governance while offering trustworthy, interoperable and secure access to sensitive data across institutional and sovereign borders where there are divergent risks, regulations and cultural norms. There is a growing need for data-governance approaches that are agile, sensitive to local values and globally scalable.[1]

The volume of sensitive data is growing, but the value is constrained. Most of it is separated and disconnected. While the rationale for keeping sensitive data in isolation is logical, as security, integrity and availability are a top priority, the separation of sensitive datasets limits the opportunity for continued innovation and discovery. The scales are not balanced. The need for protection currently outweighs the incentives for innovation.

The question arises: Are our current technical and governance systems capable of achieving a better balance between data protection and data innovation? Are our current approaches to governance equipped to address the complexity, volume, velocity and responsibility associated with sharing sensitive data on a global scale?

Data siloes and sensitive data

Generally, the reasons why sensitive data is isolated fall into two classes of concerns. There are technical concerns (different technology standards, different implementations, different systems) and there are policy and governance concerns (different legal, regulatory, institutional and normative policies). Technical reasons are generally the result of legacy (and proprietary) technology standards. Typically, data systems have been designed for a specific purpose, and historically little attention has been paid to designing them with a view towards enabling broader access to the data. Additionally, there are security, operational, business continuity and economic factors that make it challenging to open up isolated approaches for additional purposes or participants.

From a policy and governance perspective, the isolated nature of sensitive data stems from strict national, regional and local regulatory and compliance requirements. These legal requirements can prohibit such areas as large-scale transfer, access and secondary usage. Additionally, there are a variety of requirements for gaining the consent of the data subjects.

Compounding the challenges posed by interoperability and fragmented policies is the growing lack of trust on the part of the public and providers of data. Security breaches, unauthorized use of personal data and identity theft have been drivers of new legal regimes and warnings from regulators. As the 2019 Edelman Trust Barometer highlights, trust among individuals, institutions and industry is low and threatens to derail the innovative possibilities of using sensitive data.[2]

Vulnerable populations

Given that vulnerable populations may face a greater risk of adverse consequences when highly sensitive data is misused, policy-makers should identify populations that merit a higher level of protection and determine whether specific requirements should be codified within a data-protection framework. When data is collected about a person's genetic make-up, for example, this information is so unique that its knowledge or use may potentially subject them to discrimination, stigmatization or denial of medical treatment.

Privacy vs. security

Privacy and security are overlapping and complementary concepts, but they are foundationally different. Information security concerns the confidentiality, integrity and availability of information. Privacy risks may result from authorized activity that is beyond the scope of information security. Thus, protecting individuals' privacy cannot be achieved solely by securing personal data. Security involves protecting information from unauthorized access, use, disclosure, disruption, modification or destruction. Privacy, on the other hand, is concerned with managing authorized risks to individuals associated with the creation, collection, use, processing, storage, maintenance, dissemination, disclosure or disposal of personal data.

Responding to the challenge: federated systems

At its core, the central challenge in the use of sensitive data lies in striking a balance between the competing tensions of protection and innovation. The need to protect data in a way that upholds local norms, values and regulations while also enabling innovation at the global scale is a foundational challenge and requires both technological and governance-related interventions.

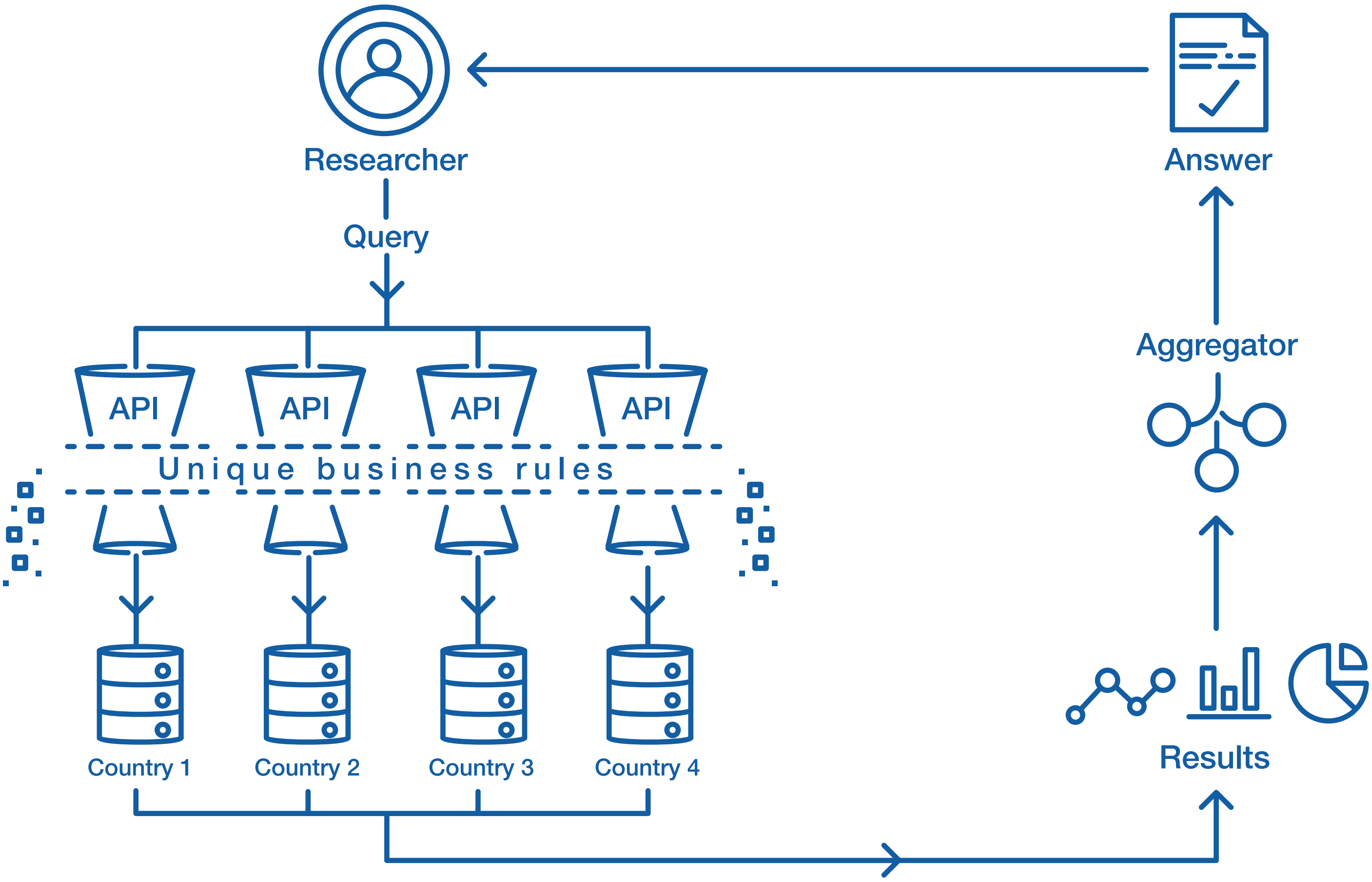
Sensitive data

Identifying and setting specific rules for categories of sensitive data is now a core part of nearly every data-protection framework. More than 75 countries define sensitive data or classify special categories of personal data. These categorizations recognize that access to different data elements (or categories of personal data) present different levels of risks to individuals. What is considered sensitive is often subjective and may vary from country to country, based on cultural, historical and economic factors. Understanding the sensitivity of a data element, or given category, is important not only in the context of protecting privacy but also for data security, information governance and risk management more generally.

One approach in balancing the need for local autonomy with global innovation at scale is the use of federated data systems. Central to this approach is federated system architecture, with multiple interconnected nodes that align on shared principles and open standards to ensure security, interoperability and reliably high performance. While physical (and legal) entities are distributed around the world, they are logically connected through interfaces that allow seamless, authorized access to secure data.

From a technology perspective, the decentralized architecture of federated systems is enabled by application programming interfaces (APIs) that provide a secure, open and reliable means of accessing disparate legacy technology systems and data formats. Along with the use of APIs, federated systems share a common, foundational technology architecture that is designed to provide functions such as differential privacy, security, identity-based access, authentication and system auditing.

Federating data using APIs



Another unique technical differentiator of federated systems is that the computation moves to the data (i.e. the data does not leave the organization). As such, the business, legal, technical and societal risks inherent with data transfer and/or centralization are greatly reduced.

Along with reducing the risks of data transfer, the decentralized architecture of federated systems can reduce operational support costs by using a shared infrastructure. Access to data via APIs reduces the amount of bandwidth required to transfer data across institutional and geographic boundaries; it also enables permission-based access with different layers of granularity.

What is an API?

- An API is an application programming interface, which simplifies the ability to retrieve data from many types of databases and applications, including those at remote locations.
- To use a non-technical example to explain how an API works, imagine you are ordering food in a restaurant. You have a menu in front of you and decide to order the special of the day. You need a way to communicate your order to the kitchen and then receive your food at your table. In this case, your waiter functions as an "API" – by taking your order, delivering it to the chefs, then bringing the food to your table when it is ready.
- While doing so, your waiter also performs other functions of an API – by ensuring clear communication (e.g. making sure the kitchen has your order and is clear about any specific requests you have made) and ensuring the quality of your meal (e.g. is the meat cooked to your specification? Do you have all of the correct side dishes?)

From a governance perspective, federated data systems can enable local control and autonomy. As such, governance can reside with local entities. This allows for a more nuanced way to address the contextual dependencies surrounding data access required by modern data policies.

Due to its highly granular, sensitive and revealing nature at the individual, family and community level, genomic data can serve as a good use-case for examining the value of federated data systems. Such data is the subject of an emerging body of effective methodologies for structuring data access, common processes for consent management, interoperable networking technologies and the global sharing of legal agreements.[3]

An example of an operational federated data system used for genomics and health research is CanDIG,[4] deployed to analyse health data across Canada. As noted on its website, "CanDIG is built to be completely distributed. Each data provider handles their own data and users, with complete control over who can access each dataset and how much, with federated analysis built on top of APIs to this data." Designed so that sensitive data can be analysed without being copied, CanDIG prioritizes privacy. From the very beginning of the project, differential privacy capabilities, advanced authorization/authentication technologies and secured API-based access were all built into the platform so that data can be analysed without exposing sensitive individual data.[5]

Another example of a federated data system in operation is the Open Algorithms (OPAL) project, an open-sourced platform designed as a decentralized means of accessing private-sector data. Because of its federated architecture and governance model, OPAL is designed to ethically extract insights from an array of private-sector data holders (such as mobile phone operators, banks, retailers and energy providers). With this approach, the data stays within the premises of the private company and only an authorized set of third-party queries (the open algorithms) have access to the pseudonymized data. What results is a safe question-and-answer system where the questions are validated in advance by a board of advisers comprised of experts and local members of the community.

With OPAL, policy-makers at the global, national and local level can create new indices regarding population density, poverty and other categories to enhance existing statistical models. Such data-querying approaches are also being evaluated to help address epidemics, poverty, financial inclusion, crime, traffic, air pollution and more. Deployment in Colombia and Senegal began in mid-2017.

Elements of a federated data system

- A federated data system allows authorized users to perform queries on the data within a federated network of organizations. The results retrieved from each organization in the federation are then aggregated and returned to the individual who submitted the query. The data never leaves the organization that holds it. Instead, the data is "visited" and only the computed answers to the query are brought back to the federation system.
- Federated data systems use foundational, shared technology architectures, including operational components of security, auditing, authentication and access rights, among others. Agreement on which functions of this architecture are shared and which are left to local control is a critical component in setting up the federation that will allow access to the data.
- A central component of federated data systems is the use of APIs, which are managed using this shared technology architecture. The use of APIs and the foundational architecture enables a scalable, secure and reliable means of accessing the local data stores of the federated organizations, even though they likely use a variety of technology systems and data formats.
- Most importantly, the use of APIs allows the definition and enforcement of specific governance policies (including honouring local laws) by each organization within the federation. The use of APIs within a federated data system allows for crucial governance control to reside with each local entity in the federation, based on the overall agreement of the federation.

The office of European Statistics is also evaluating the use of federated data systems for its work on Trusted Smart Statistics for National Statistical Institutes. The intent of this work is to establish a set of policy indicators that would be informed by an array of private-sector inputs from multiple industry sectors. As the European Commission notes, “Statisticians have a new opportunity to produce official statistics that use an extended data ecosystem. These Trusted Smart Statistics are a service provided by smart systems, embedding auditable and transparent data life cycles, ensuring the validity and accuracy of the outputs, respecting data subjects’ privacy and protecting confidentiality.”[6]

The benefits and constraints of federated data systems

Benefits

- Enable local control with global scale and efficiencies
- Addresses both privacy and security concerns
- Facilitate the ability to discover new data
- Enable the ability to analyse larger datasets to gain richer insights
- Reduce financial and operational costs
- Facilitate cross-border data sharing by respecting local governance and legal regulations

Constraints

- Add extra complexity to decision-making processes
- Create new types of “infomediaries” where the risks and liabilities are unknown
- Redress and remediation measures for accidents or acts of negligence with federated data are not yet developed

Emerging policy and governance uncertainties

While the adoption of federated systems is gaining momentum, there are many policy and governance-related uncertainties. One of the most significant challenges in this regard is gaining the agreement and commitment from federation members on the shared goals, operating principles, metrics for success and apportionment of the benefits and commercial value. Gaining alignment among the various stakeholders – including beneficiaries and vulnerable populations – on the goals and intended outcomes of the federation is a significant and time-consuming challenge.

From a technical viewpoint, there is a need for standards-based, open and widely available hardware and software components on which to build the platform. This platform should be highly scalable and able to perform many of the fundamental functions required for the use of APIs and for security, authentication, auditing and access rights.

Data access

Even when an open, standards-based platform is available, a core challenge will be defining who will have access to what types of data, under what circumstances and for what purpose. When defining the custodianship and approval processes to authorize data access, is it a collective decision-making process, in which each federation member has a voice? Or can federation members act unilaterally? Are there various "levels" of data access that depend on the functional role of the person making the query? Are some federation members given universal rights to query all levels of data, while others are limited to only a subset? Do government and academic researchers have the same access rights as those from the private sector?

Threshold of entry and exit

Another concern is the threshold of entry for new members to join the federation. Questions must be addressed regarding the provenance, permissions, value and quantity of data types they are contributing, and whether the data can be queried without individual consent or if the member organization agrees to use only the derived insight in ways stipulated by the federation. Questions about the process to exit a federation are also relevant, including when and whether to recontact the entity or the data providers that are no longer part of the federation. Can the data from that entity be withdrawn from use?

The unique challenges of sharing genomic data

The sharing of genomic data can be challenging as it is the most unique data about a person that exists. A full genome sequence is an unalterable depiction of an individual and much more inimitable than biometric information or any other personal identifier.

When an individual shares their genomic sequence data, they are also sharing a huge amount of information about their parents, siblings and ancestors, since genes are passed on through subsequent generations. Recently, "cold case" murders in the US have been solved through the use of genomic "data sleuthing". Authorities used decades-old DNA found at crime scenes, and public databases of genomic information were then able to follow genetic clues to define a very small pool of possible suspects, which eventually led to a definitive match. These cases have caused heightened privacy concerns, as the databases did not include information on the suspect. They contained genomic data on one or more of the suspect's relatives, but this was enough to allow accurate identification of someone whose data was not held in the database.

In healthcare, anonymized data is shared consistently without fear of reidentification. Due to the unique nature of an individual's full genome sequence, however, it may become possible to reidentify that person if enough additional identifiers can be matched with the genomic information. While unlikely, there is a greater chance of reidentification based on sequenced genomic information.

It is important to note that the above issues are related to the sharing of the actual genomic sequence data, not the results of individual genomic tests. If test results (e.g. genes showing mutations) are anonymized, they do not generally pose a heightened risk of reidentification.

Financial model

The initial funding and ongoing economics required to create and maintain a federated system are critical to ensuring its long-term impact. The economic model that underlies the federation can raise many questions. Will the federation allow the data (and the models derived from its use) to be used for commercial purposes? Will the system be funded by government, subsidized by donors, commercialized by freemium subscription models or delivered as an ongoing service with defined commercial terms? The question about how these various economic models will align with the federation's principles and how they will evolve over time must be addressed upfront and in a transparent and inclusive manner. While the federation's potential to create significant socioeconomic value through access to distributed sensitive datasets may be quite clear, potential federation members will need a compelling business case to engage as a stakeholder on a long-term basis.

Transparency

Enabling meaningful engagement of individuals in the design of federated systems is one way to anticipate and plan for the impact (both risks and benefits) of data access and use to individuals and populations. Involving the data providers and representatives of populations who will be affected by access to and use of their data, and any breaches of security or other operational norms, is critical to understanding expectations, red lines and acceptable trade-offs during system design.

One approach to enable meaningful transparency regarding data and data access is the notion of dynamic consent, which is gaining attention. Rather than having patients consent to the use of their data in a binary, all-or-nothing manner, dynamic consent allows individuals to change their mind about the amount of data they wish to be used, and for what purpose. This provides a transparent and ethical vehicle allowing individuals to modify or withdraw their consent if they change their mind later. For researchers or other "data managers", dynamic consent allows for better electronic tracking of consent, including the details of that consent. It is also hoped that people will be more inclined to consent to their data being used if they are given more flexibility to change their mind at a later date.

Data ownership

Data ownership within the context of a data federation can be complex, with a variety of legal interests to consider across jurisdictions. Federated systems require open discussion on the sensitive datasets required to achieve their intended goals. Given the diversity of commercial, regulatory and social considerations that entities within a federation have to navigate, stakeholders will have different concerns and competing incentives with regards to how, if, when and by whom it can be used.

Conclusion and next steps

The technical, commercial, social and regulatory challenges related to setting up and governing federated data systems are clearly complex. Federated data systems may provide a means of addressing that complexity and delivering value by allowing data to globally interconnect in ways that balance the potential risks and that address local privacy norms and data-protection regulations. Because there are few case studies available on fully functional federated systems, we need to continue to test whether such systems allow for data to be accessed across borders.

Continued progress will require leaders to focus on important items identified in this white paper. Some of the most critical areas include:

Development of open technical standards

To accelerate the adoption of federated systems, an open source and widely available technology platform on which to build federated data systems needs to be advanced. This platform needs to allow for various policy-enabling and privacy-enhancing technologies (well-designed APIs, anonymization) to strengthen the responsible use of data. Work on data and metadata standards should be recognized and encouraged by the organizations that use and benefit from them, including academia, industry, government regulators and funding agencies.

Development of robust and agile governance frameworks

Leaders need to support the development of real-world pilots that establish governance frameworks for federated data systems. These approaches should balance the dynamics between data risk and benefit as well as the importance of local culture, values and social norms. Leaders should look to sectors such as precision medicine (which has been working on this issue for many years) for techniques and knowledge exchange. An example is the Global Alliance for Genomics and Health's (GA4GH) Framework for Responsible Sharing of Genomic and Health-Related Data. It enables transparent, responsible and accountable insight sharing. Many pilots around the world are using this framework and its international data-sharing code of conduct as a way to innovate in the use of sensitive data while enhancing privacy and security in a proportionate manner.

Strengthen stakeholder trust

Engaging stakeholders in a structured, robust awareness-building campaign on the value of federated systems will be helpful in strengthening trust in their ability to ethically share data and create accountability and sustainable socioeconomic value. Dialogue must take place among all the stakeholders, in order to resolve a host of underlying tensions. New methods of encrypting and sharing genomic data in a way that enables collaborative research without compromising patient privacy are needed – and could help to build trust with those who are providing sensitive data.

Accelerate pilots that deploy federated data systems

The design and testing of federated data systems can be accelerated by "exercising" their use through agile, lightweight experimentation. Projects can be piloted in "sandboxes" or other safe spaces, allowing experimentation with governance frameworks, technologies and consent processes, to ensure systems can be developed that encourage transparency, security and trust.

Focus on shared principles

Human-focused design, based on simplicity, transparency, efficiency and ease of use, must lie at the heart of the data practices, permissions and controls provided within federated data systems. These systems need to build simple-to-use tools that encourage individuals to become engaged in creating data-use policies and allow them to change those settings over time without facing undue penalties. Providing users with these capacities is critical if they are to understand and find value in the shared benefits and outcomes from data access in a federated system. Additionally, federated systems are unique in their ability to enable fundamentally new approaches to governance that can create rules that are both robust enough to be enforceable and flexible enough to accommodate contextual differences.

Federated data systems are a critical component of powering a sustainable, inclusive and transparent digital economy. We believe that the widespread use of federated data systems can stimulate innovation by allowing access to sensitive data while minimizing the risks and unintended consequences. By raising appropriate questions that will help guide policy-makers and leaders, the hope is that this white paper highlights important issues – and ways to address them – when designing federated systems for accessing sensitive data in a scalable, trustworthy and legally compliant manner.

Proactive engagement from government officials, business leaders and civil society members is required to establish real-world pilots and enable an iterative process of multistakeholder dialogue, piloting, experience-sharing and refinement. The World Economic Forum is ideally suited to support this work as part of its mission to improve the state of the world. The Forum's Breaking Barriers to Health Data project explicitly brings together a coalition of partners to design and test governance frameworks that enable federated data systems to organize and operate, with the goal of enabling cross-border access to sensitive health data.

Acknowledgements

Lead author

Arthur Hermann, Kaiser Permanente Fellow,
Centre for the Fourth Industrial Revolution,
San Francisco, USA

With thanks for additional comments and insight from the World Economic Forum's Data Policy platform team.

Endnotes

1. World Economic Forum, Data Policy in the Fourth Industrial Revolution: Insights on Personal Data (11 November 2018), https://www.weforum.org/whitepapers/data-policy-in-the-fourth-industrial-revolution-insights-on-personal-data (link as of 15/7/2019).
2. 2019 Edelman Trust Barometer: Global Report, https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_Trust_Barometer_Global_Report.pdf (link as of 15/7/2019).
3. Global Alliance for Genomics and Health (GA4GH) Framework for Responsible Sharing of Genomic and Health-Related Data (2014), https://www.ga4gh.org/wp-content/uploads/Framework-Version-10September2014.pdf (link as of 15/7/2019).
4. CanDIG (2019), https://candig.github.io/ (link as of 15/07/2019).
5. Ibid.
6. Fabio Ricciato (2018), European Commission, EUROSTAT, Towards a Reference Architecture for Trusted Smart Statistics, https://ec.europa.eu/eurostat/cros/system/files/dgins2018_tss_ricciato.pdf (link as of 15/7/2019).

WORLD ECONOMIC FORUM

COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744

contact@weforum.org
www.weforum.org