Trusted Data Space for Creating Value from Data in a Chain Reaction Manner

Tomoaki Washio, Hiroki Itoh, Koki Mitani, Gembu Morohashi, Kenji Umakoshi, Tetsuya Okuda, Kazuyuki Takaya, Kei Ohmura, and Gen Takahashi

Abstract

The Smart World holds the possibility of maximizing the value of data throughout society by enabling diverse organizations to bring each other data for analysis and create data for new purposes in a chain reaction manner beyond the walls of companies and industries. In reality, however, the use of data among organizations goes no further than the provision of limited data to limited parties, which prevents a value chain from being achieved. This article introduces a new mechanism for data sharing called "trusted data space" as an initiative to solve this problem and describes key technologies for making it a reality.

Keywords: security, data sharing and utilization, data space

1. Smart World enabling data sharing across industries

There has been much activity in the research and development of technology for not only reproducing real-world systems such as manufacturing lines and chemical plants in cyberspace to analyze and predict system operations but also for data sharing and analysis beyond individual organizations, business fields, and industries.

In a smart city, for example, a massive amount of output from sensors, video cameras, etc. installed in physical space can be converted to data so that the movements of people and things can be analyzed in cyberspace in a cross-sectoral fashion. The results of this analysis could then be used as a basis for directing the behavior of people and things in the city, that is, in the physical space. The Smart World that fuses physical space and cyberspace in such an advanced manner is fast approaching.

In the Smart World, data will be continuously generated from a variety of individuals and companies on a global scale at a level of quality and volume not seen before. There will therefore be a need for a mechanism that can effectively use this massive amount of diverse data. In particular, there is a need for a data marketplace in which everyone can bring each other data for analysis and generate data for new purposes in a chain reaction manner beyond the walls of companies and industries. This would have the effect of uncovering value from each other's data and maximizing the value of data throughout society. This mechanism is called "trusted data space" (hereafter, data space).

Both providers and users participate in a data space. The data in a data space, while placed under the management of the providers, can be used virtually in the manner of one massive data lake and searched through freely. The providers present terms of data use (period of use, allowed processing, secondary use conditions, etc.), while the users can use the data within the allowed range after agreeing with those terms. This type of mechanism enables data collected for a certain purpose to be used for a new purpose,

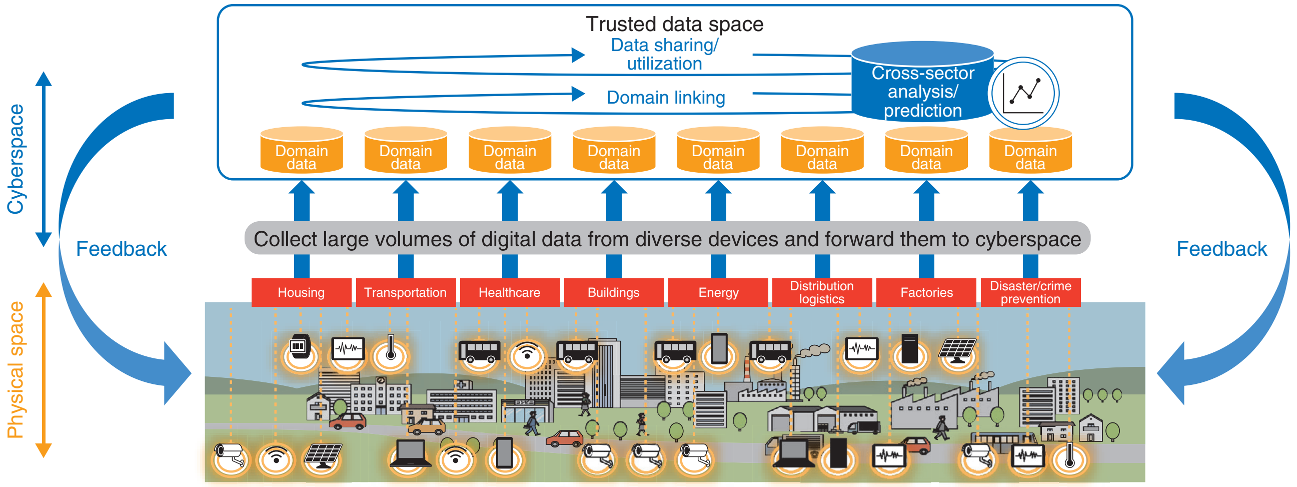

Trusted data space
Data sharing/ utilization
Domain linking
Cross-sector analysis/ prediction
Domain data
Cyberspace
Physical space
Feedback
Feedback
Collect large volumes of digital data from diverse devices and forward them to cyberspace
Housing
Transportation
Healthcare
Buildings
Energy
Distribution logistics
Factories
Disaster/crime prevention

Fig. 1. Smart World and trusted data space.

which points to the possibility of generating data with new value from existing data in a chain reaction manner. For example, marketing data collected through the cooperation of amusement facilities, transportation operators, and eating and drinking establishments for making user recommendations on personal behavior on a certain day could also be used by local governments for various purposes such as disaster and crime prevention, promoting the health of its residents, and urban development. In short, things that could not be achieved by a particular business entity alone would become possible (**Fig. 1**).

1.1 Issues with cross-organizational use of data

Despite heightened expectations for such a society, the social adoption of data sharing and analysis technology remains at the level of local value discovery. Data collected by an organization is generally used only for the purpose for which the data were collected by the collecting company, while the use of data among organizations stops at the provision of limited data to limited parties, preventing the creation of a value chain. We call this the “data sharing wall.”

Overcoming the data sharing wall requires technology for resolving three types of issues: issues in discovering an appropriate data provider and optimal data, issues in forming an agreement on data use, and issues in sharing and using data on the basis of that agreement (**Table 1**).

By providing technology for resolving these issues, we should be able to achieve a form of data sharing that creates new value in a chain reaction manner while also maximizing the value of data throughout society.

2. Trends in data sharing throughout the world

In Europe, data sharing is being revitalized with a focus on the manufacturing industry. Typical of this movement is the Gaia-X [1] project that aims to establish a data-sharing infrastructure for Europe. The Gaia-X vision of supporting data sharing and use on a European scale was announced on October 29, 2019 by the German and French governments. This was followed by the founding of the Gaia-X AISBL non-profit organization for achieving this vision in January 2021. The plan is to construct an infrastructure that can provide a technical mechanism for ensuring interoperability with diverse cloud services while controlling data access based on rules and agreements and protecting data sovereignty*. In Japan, the Data Society Alliance was founded on April 1, 2021 in the wake of this movement with the aim of constructing a platform called DATA-EX [2] to facilitate data linking across diverse fields. A number of projects have been established as data sharing initiatives using Gaia-X with the aim of constructing data-sharing infrastructures composed of companies in a trustworthy relationship such as a supply chain. In Germany, for example, there is Mobility Data Space for achieving Mobility as a Service (MaaS) and

* Data sovereignty: The right of a data provider to determine the range of data disclosure, usage applications, etc.

Table 1. Issues with cross-organizational data sharing.

Issues in discovering appropriate data providers and optimal data	• Prevent data giving/receiving among undesirable parties • Enable data users to discover data applicable to their conditions
Issues in forming an agreement	Form an agreement between the data provider and data user on data-processing conditions and process the data on the basis of those conditions
Issues in sharing and using data on the basis of the agreement formed	Enable the data possessed by a provider to be processed and provided to data users without disclosing the data and processing method to the other party

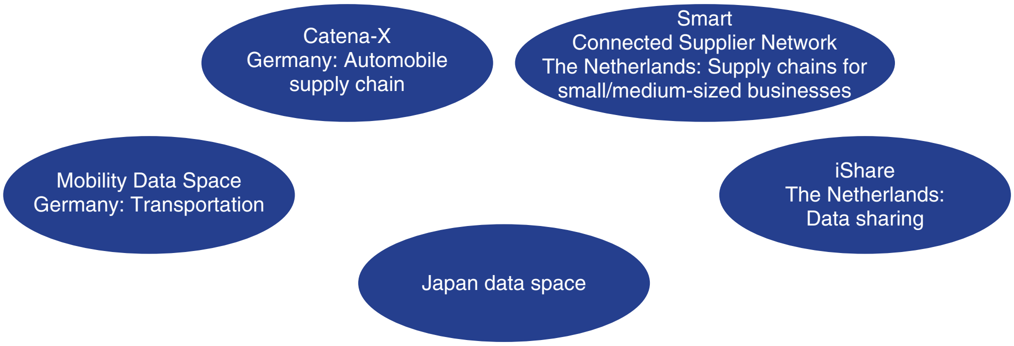

Catena-X
Germany: Automobile supply chain
Smart Connected Supplier Network
The Netherlands: Supply chains for small/medium-sized businesses
Mobility Data Space
Germany: Transportation
iShare
The Netherlands: Data sharing
Japan data space

Fig. 2. Data sharing platform beyond industry walls.

Catena-X for achieving an automobile manufacturing supply chain, while in The Netherlands, there is the Smart Connected Supplier Network (SCSN) for small/medium-sized businesses in the manufacturing industry. Catena-X and SCSN plan to build data sharing infrastructures on a scale of 1000 and 3000 companies, respectively, by 2022.

The construction of data-sharing infrastructures across industries has already begun. However, this process is still at the stage of setting up rules centered about Gaia-X; as a result, data sharing has been limited to companies that are already in a trustworthy relationship. In addition, major issues are expected to arise from here on in terms of providing security to protect data from cyber attacks and providing protection of data sovereignty that enables data sharing while protecting the rights of data providers. The questions to be answered are what technologies should be implemented to achieve these goals and how should such a system be constructed as a social infrastructure (**Fig. 2**).

3. Technology for configuring a data space

The following three mechanisms must be considered to resolve the issues described in Section 1.

(1) Mechanism for discovering data, applications, and business partners that can be trusted: Catalog data/applications and business partners and visualize information on their reliability to gauge their appropriateness. Match up with business partners that can be trusted on the basis of this visualized information.

(2) Mechanism for forming an agreement on data processing conditions: Form an agreement between the data provider and data user on data-processing conditions, disclosure conditions, and processing methods on the basis of cataloged information. Determine whether a certain type of data processing can proceed according to the content of that agreement.

(3) Mechanism for processing data based on an agreement: The data provider manages the data on its own and virtually shares and integrates the data only when necessary. The data user executes various types of analysis and processing deemed necessary while keeping secret not only the data but the data-processing methods as well.

Current data processing executes risk management on the basis of contracts such as non-disclosure agreements (NDAs) and operation management policies, but there is a need for corroboration based on technology, as in mechanism (3). To execute transactions with unknown parties with which no business relationship has been established and create value in a chain reaction manner, there is a need for forming an agreement on data-processing conditions, as in mechanism (2). Finally, there is also a need for discovering partners, data, and applications that can be trusted, as in mechanism (1).

We introduce a data-processing mechanism based on an agreement centered about a data space. This mechanism gathers and virtually integrates data dispersed among organizations and individuals in an encrypted state through the use of a virtual data lake and processes that data while they are encrypted by using either of two methods: data sandbox or secure computation. We also introduce our forward-looking work on an agreement-forming mechanism for data-processing conditions.

3.1 Virtual data lake

When sharing one's data with another party, only the minimum required amount of data must be provided, and information protection and management must be executed according to data usage conditions established beforehand. A virtual data lake is achieved through technology that virtually integrates data under different management entities scattered over a wide area while maintaining the governance of each management entity and that transfers only a minimum amount of data on the basis of the requirements of the data user. On top of this, a virtual data lake includes mechanisms for quickly informing the data user that data generation has begun and for allowing the data user to begin using a portion of that data before data generation is fully completed. This makes it possible to use even a large amount of data at an early stage as needed and accelerate data sharing.

For example, a company that is generating data could simultaneously generate and update data management information (metadata such as a data catalog and control policies) and notify the platform of such. The platform, in turn, could appropriately control the sharing of that data by informing approved users of the existence of that data and transmit to them that portion of the data deemed necessary on the basis of data management information. In addition, users would be able to find desirable data with good efficiency by referring to a virtually integrated list of data and begin using those data at an early stage. These mechanisms enable safe and convenient use of data beyond individual organizations.

3.2 Data sandbox technology

Data sandbox technology brings together an organization that possesses data but no analytical technology and that has no desire to share that data with another company and an organization that possesses analytics technology but has no desire to share that technology. Therefore, analysis results can be obtained without these organizations having to share data and analytics technology with each other.

If a data owner and analytics technology owner were to reach an agreement on the use of a data sandbox, a dedicated data sandbox that includes data and analytics technology could then be created on the basis of that agreement. A data sandbox blocks communications with the outside and encrypts internal communications, storage, and memory so that even the data sandbox provider cannot decrypt that content. In short, data can be analyzed without anyone including the data owner, analytics technology owner, and data sandbox provider accessing decrypted data or analytics technology.

The data sandbox saves the results of analysis in a location that can be viewed by the data owner and/or analytics technology owner on the basis of the agreement reached between those two parties. First, the case in which the data owner may view the results of analysis means that the data owner can access analysis results without having to share data with another party and that the analytics technology owner can be compensated for the use of its data analysis technology by the other party without having to share algorithms with another party, all through the use of a data sandbox. Next, the case in which the analytics technology owner may view the results of analysis means that the data owner can be compensated for the use of its data for analysis by the other party without having to share data with another party and that the analytics technology owner can access analysis results using another party's data that it does not possess without having to share analytics technology with another party, all through the use of a data sandbox.

3.3 Secure computation

Secure computation is an advanced encryption technology that can process data while they are encrypted without returning to the original data even once. In 2019, NTT developed secure computation deep learning to provide security and privacy measures for data used in artificial intelligence (AI). This technology executes deep-learning training and prediction while keeping the target data encrypted without returning to the original data even once. Since the conventional technology had performance issues, this technology became an alternative using processing that was even simpler than that of ordinary (unencrypted) deep learning training and prediction. NTT's secure computation deep learning uses world-class secure computation-processing performance and reproduces the training process using standard optimized processing executed in deep learning as a world's first by secure computation.

In other words, all the steps required for data usage in deep learning, that is, (1) data provision, (2) data storage, (3) training, and (4) prediction, can be executed in an encrypted state. Since data are always kept in an encrypted state without returning to their original form even once, this technology enables users and organizations to provide data with peace of mind compared with conventional technology. This should lead to an increase in the amount and types of data that can be used for training. It is exactly this expansion of data that should enable AI to achieve even more accurate and advanced analysis.

3.4 Agreement-forming mechanism

In the use of data, there are two conditions under which the provider approves of data usage and requirements that the user demands of that data. The content of those conditions and requirements differ depending on the provider and user. The provider may specify as conditions under which parties are permitted to use the data, the purpose of use, the range of use, the period of use, etc. The user, on the other hand, may specify as requirements the target data, purpose of use, desired processing, etc.

To enable data usage in a form that both sides agree upon, there is a need for a mechanism that can compare data conditions and requirements of data usage between the provider and user then form an agreement.

It has been common for a provider and user to express provision conditions and usage conditions, respectively, in the form of policies, which would then be checked manually by each other to form an agreement. In the future, however, we can consider an approach in which the content of a user request for data usage is compared with those policies at the time of that request to automatically determine whether the agreement is being satisfied. If differences exist between requirements and conditions, the ability to dynamically adjust those requirements and conditions between the provider and user would enable more flexible agreement formation. This kind of approach must also be targeted for future study to promote the use of data among companies and organizations.

4. Toward the future

The creation of a data space should accelerate the sharing of data beyond corporate and industry walls, which has been difficult, and enable data sharing that creates new value in a chain reaction manner. To make the data space a reality, we plan to research and develop key technologies while accelerating the testing of those technologies with partners.

References

[1] Gaia-X, https://www.gaia-x.eu/

[2] DATA-EX, https://data-society-alliance.org/about/vision-mission/#top

Tomoaki Washio

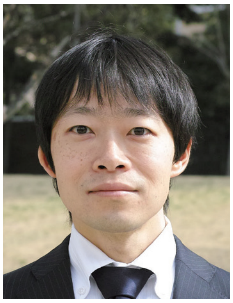
Senior Research Engineer, Social Information Sharing Research Project, NTT Social Informatics Laboratories.

He received a B.E. and M.E. in systems and information engineering from Hokkaido University in 2000 and 2002. Since joining NTT in 2002, he has been engaged in research and development (R&D) of authentication systems and secure data sharing technologies.

Hiroki Itoh

Senior Research Engineer, Social Innovation Research Project, NTT Social Informatics Laboratories.

He received a B.S. from Tokyo University of Science in 2002, M.E. from Tokyo Institute of Technology in 2004, and M.S. in management of technology from Tokyo University of Science in 2009. He is engaged in leading practical application of research outcomes from NTT Social Informatics Laboratories.



Koki Mitani

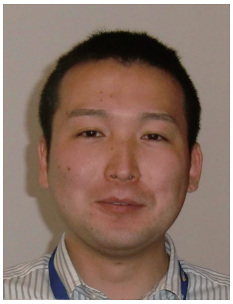
Senior Research Engineer, NTT Social Informatics Laboratories.

He received a B.E. in information and computer science and M.Sc. in engineering in science for open and environmental systems from Keio University, Kanagawa, in 2003 and 2005. In 2005 he joined NTT. From 2011 to 2015, he was a product manager of global network services in NTT Europe Ltd. and NTT Communications Corporation. He returned to NTT in 2015, where he currently leads open and collaborative innovation for building global infrastructure for data sharing across businesses at NTT Social Informatics Laboratories.

Gembu Morohashi

Senior Research Engineer, Social Information Sharing Research Project, NTT Social Informatics Laboratories.

He received a B.S., M.S., and Ph.D. from the University of Electro-Communications, Tokyo, in 2001, 2003 and 2009. He began working at NTT in 2003, and his main research interests are cryptography and information security.



Kenji Umakoshi

Senior Research Engineer, Social Information Sharing Research Project, NTT Social Informatics Laboratories.

He received an M.E. from Waseda University, Tokyo, in 2009. He joined NTT in 2009 and has been researching and developing in areas such as ubiquitous/Internet of Things computing, smart room/factory, and data sharing platform. He also worked as a product manager of a cloud service in NTT Communications Corporation from 2014 to 2016.

Tetsuya Okuda

Research Engineer, NTT Social Informatics Laboratories.

He received a B.S. and M.S. from the University of Tokyo in 2009 and 2011. Since 2011, he has been engaged in R&D on security protocol at NTT. He is a member of the Information Processing Society of Japan (IPSJ) and received the IPSJ/Computer Security Symposium Student Paper Award in 2019.

Kazuyuki Takaya

Senior Research Engineer, Supervisor of Data Sharing Infrastructure Project, NTT Software Innovation Center.

He received an M.E. from Waseda University, Tokyo, and joined NTT in 2000. He is engaged in R&D of technologies and platforms for data sharing.

Kei Ohmura

Senior Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

He received an M.E. from Waseda University, Tokyo, in 2009. Since joining NTT the same year, he has been engaged in developing platforms for cloud, Internet of Things, and AI leveraging open source software.

Gen Takahashi

Senior Research Engineer, Social Information Sharing Research Project, NTT Social Informatics Laboratories.

He received a Master of Media and Governance from Keio University, Tokyo, in 2005 and joined NTT in 2006. His research interests include information security and cryptographic engineering.