No.49 Apr.2014 ISSN 1883-1974 National Institute of Informatics News

NII Today

National Institute of Informatics News

国立情報学研究所
National Institute of Informatics

FEATURED **TOPIC**

CPS
Connecting the Real World and the Cyber World

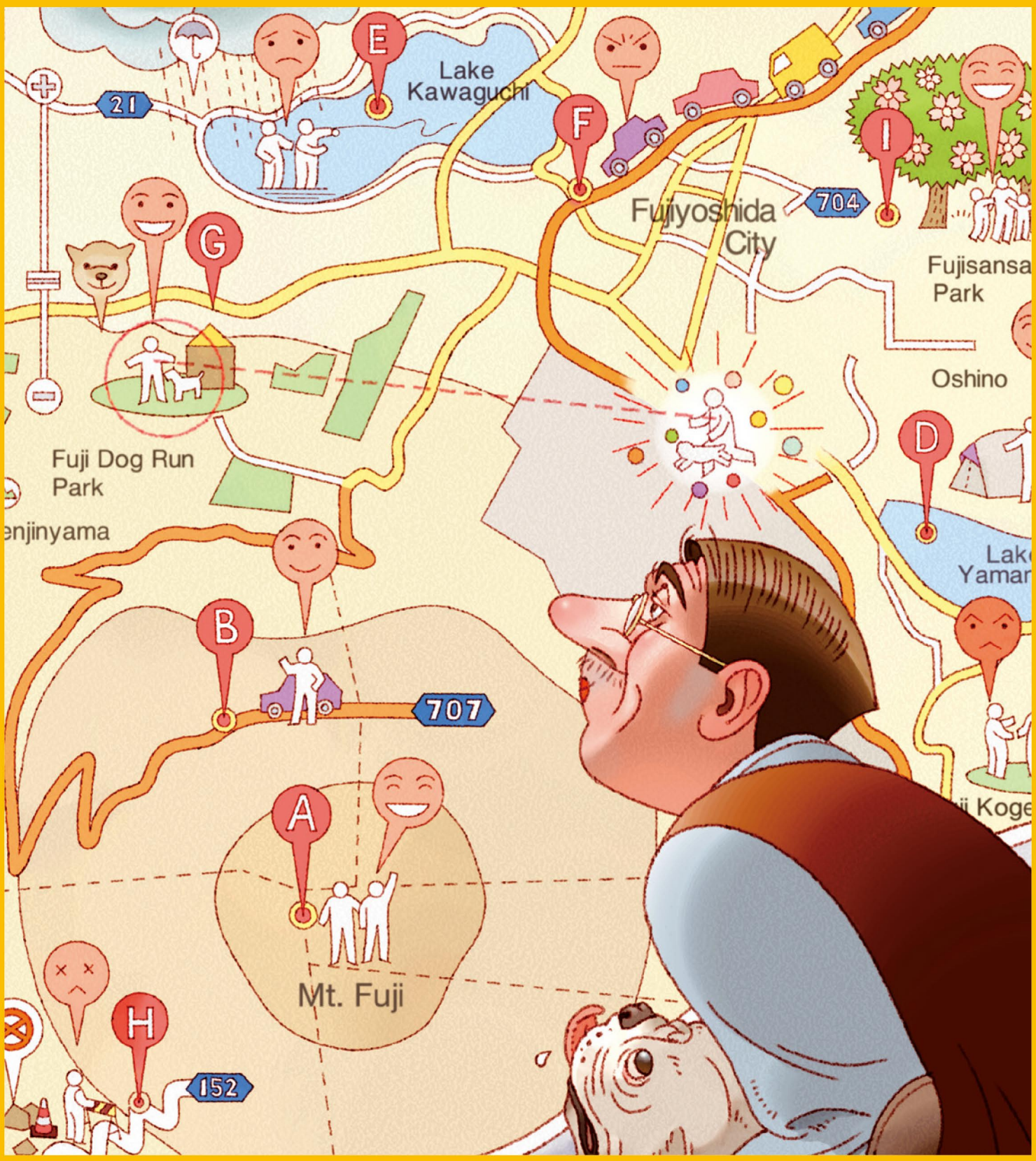
NII **Interview**
I Want to Connect the Cyber World with the Real World, and Create New Value that Enriches People's Lives

NII **Special 1**
Big Data Technology that Changes Society

That's **Collaboration 1**
Attracting Visitors to Tourist Sites by Leveraging Mobile Phone and Smartphone Data

NII **Special 2**
Creating Rules to Make Full Use of Personal Data

That's **Collaboration 2**
At the Forefront of Compression/Decompression Technologies



INTERVIEW WITH

Kenro Aihara

Associate Professor, Digital Content and Media Sciences Research Division, NII
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies

I Want to Connect the Cyber World with the Real World, and Create New Value that Enriches People's Lives

You can browse the Internet anywhere today – in a shopping mall and even on a commuter train. You can share your thoughts and experiences using services like Twitter and Facebook. No matter where you are, you are now always surrounded by the information space. A new research movement, called Cyber-Physical Systems (CPS), is underway to integrate information space even more tightly with the physical world, producing value from their interaction. NII Associate Professor Kenro Aihara, a pioneer in this field, describes the future in which the cyber world and the real world meld together.

Taki Cyber-Physical Systems (CPS) is a term that most people are not familiar with today.

Aihara The world of information science began in the form of computer science before the start of the Second World War. The development of computers was the major theme then. After information science entered the Internet era in the 1990s, research on processing information on networks, such as how to handle and share information over the Internet, took center stage. CPS is the next paradigm. It has been said that happenings in the real world are reflected on the net with services like Twitter. But you can't understand real-world phenomena with just data on the net. We have to incorporate what is happening in the real world into the data, analyze it, and from that, produce new value and bring it back into the real world. CPS is the integration of the real world and the information world, of physical space and cyber space.

Taki How is this different from the approach of connecting objects of every kind to the network – the Internet of Things?

Aihara The Internet of Things is one way of realizing CPS. But it's not enough. Even if we connect everything, we are still missing people. How they think and act, and the value they find in that, are not included. Even if we can understand who bought what and when by analyzing point-of-sales data, we only grasp what comes at the end of the human behavior called "buying." We can't understand a person's thought process that led to that final point, and it fades away. I think gaining this understanding is critical. This is called "context sensing." It means paying attention to emotions and feelings that provide the context for human behavior.

Taki Why are you so concerned about context?

Aihara There is no value in just relying on data and contents. Value is produced when something is interpreted by a person's value system – value then becomes information. For example, a piece of data stating "The highest temperature is 15°C" means completely different things to an office worker and a retail store clerk. Without grasping the context, you can't understand the value some content has for a person. Information born from circumstances is content that has value, because it includes context.

Another reason is that the value of information received by a person differs depending on the context in which he or she is placed. Is he or she energetic? Or tired? We need technologies that select contents for recipients based on their context.

Taki It's not easy to convert a person's emotional state into data, is it?

Aihara Smartphones and cars now serve as masses of sensors that record the behavior of people. They surround people and allow us to understand their behavior. To study how people spend their time in a city, my team created a specialized app and asked users to install it on their smartphone.

We currently have the cooperation of several hundred people living, working, and shopping near Futako-Tamagawa Station on the Tokyu Denentoshi Line (Setagaya Ward in

Tokyo). We are partnering with Tokyu Corporation and the local neighborhood association to provide information about events and announcements through smartphones. And we are gathering behavior logs not only on users' movements, but on what they paid attention to and where they moved as a result. We have also begun analyzing users' facial expressions when they look at their smartphone screens. We are not capturing the images of faces. Instead, our app looks at facial features and sends the data to our server for analysis so that we can understand how people's expressions change based on emotions. Our technology can deduce the moods of users -- for example, if they are happy or tired. We can understand not just where people are gathered in a city, but the qualities of the place or event that excites them.

Taki What do you seek to achieve through such an effort?

Aihara The goal of our work in an area like Futako-Tamagawa is to increase the overall value of a community. To develop a community, planners study the flow of people, but I think in general they depend a lot on their experience. I think we can provide data-based community development and answer questions like, "Where can we effectively hold an event?" I also want visitors to enjoy their time. What contents can we provide to stimulate their actions and bring delight? I think of alluring community development as an operating system. I want to install it in the social infrastructure of a community. To do this, I am exploring how we can leverage technologies like facial expression recognition.

Taki What other CPS-related research are you pursuing at NII?

Aihara The smartphone app I'm involved in is one approach. In addition, we are conducting research to discern trends in the world from data provided by mobile carriers and automobile companies. Many people have heard about GPS navigation providing guidance on which open roads to take in the aftermath of the Great East Japan Earthquake. From GPS data of car movements, we can understand the nature of a traffic congestion – for example, if it occurred naturally or because of an accident. By integrating such data with weather information, news images, and Twitter postings, we begin to get an objective view of a situation. Users can optimize their actions based on these kinds of circumstances.

These days, people tend to flock to popular tourist spots because of what Internet search results tell them. Top-ranked spots become crowded, and other sites don't draw visitors, all because people base their behavior on search advice. If we can provide the right information, tourists can enjoy the perfect spots for them without dealing with crowds. Right now, I'm advising tourism bureaus on projects that study tourist movements.

> **Information born from circumstances is content that has value, because it includes context.**

A Word from the Interviewer

Near the end of last year, Professor Yoh'ichi Tohkura, former NII deputy director general, passed away. Professor Tohkura stressed the importance of the interface between human beings and products and technology, and he pioneered much of the research in this field. He taught journalists a lot about the future of information communication technology, from the time he served as research director at NTT and ATR, when "multimedia" was a newly coined term. Even now I can remember his animated face when he talked about new technologies.

Associate Professor Aihara says, "Professor Tohkura gave me both emotional and practical support. The CPS project is possible thanks to his backing." Associate Professor Aihara said that when he was once anxious that he could not continue the project with only external funding, and could not hire researchers, Professor Tohkura came to the rescue. Rest in peace, Professor Tohkura.

Junichi Taki

Editorial Writer and Senior Writer, Nikkei Inc.

Joined Nihon Keizai Shimbun (Nikkei) after graduating from the School of Political Science and Economics, Waseda University. Held positions in the Industry Department and Washington DC News Bureau, and was the Senior Staff Writer in the Economics Department at Osaka Head Office and Head of the Science and Technology News Department at Tokyo Head Office amongst other posts, before taking up his present position in March, 2009. He is in charge of science and technology, environment and medicine.

Big Data Technology That Changes Society

CPS Applications and Challenges from Integrating Diverse Data

Extracting a mere portion of information from social system operations and people's activities in their communities results in massive data on the petabyte and exabyte scale. What's more, the data's formats are diverse, and it is difficult to assimilate them into a single database. How should big data from sensor networks, SNS, and other sources be collected, stored, and analyzed to provide valuable feedback to the real world? Professor Kazuo Imai of NII, a respected authority on ubiquitous computing, and Professor Atsuhiro Takasu, an expert on data analysis, share their views.

Kazuo Imai
Professor by Special Appointment,
Research Strategy Office, NII

How will CPS change the real world?

—How will CPS (Cyber-Physical Systems) change the real world we're living in? Please tell us about a specific scenario.

Imai An easy-to-understand example is a field test Hokkaido University is planning as part of our CPS research along with Osaka University and Kyushu University. The "Smart Snow Removal Field Test" seeks to implement optimal snow plow operations to relieve traffic congestion caused by snowfall. We need a variety of data to achieve this goal. To understand snowfall conditions, we need, of course, meteorological data. But it is also effective to have data from instruments measuring road conditions and text information from pedestrians using social media. To understand traffic conditions and the progress of snow removal, we need real-time information. Besides taxi movement data, we also want to get operational data from snow plow trucks – from their drivers' smartphones, for example. What's more, in addition to the real-time data, we also need to convert past records of traffic congestion and snow removal operations, traffic accident reports, and geographic data into practical knowledge. By gathering all the data and processing it with analytical, simulation, and visualization techniques, we can produce results that lead to instructions (decision-making) on the most effective ways to remove snow. Achieving this will allow buses and emergency vehicles to travel smoothly, and improve the average speed of traffic on roads that tend to get congested. We also expect such a system to help prevent accidents by providing information on frozen road surfaces.

If we can skillfully analyze massive and complex data, a wealth of applications becomes possible. For example, we can offer systems that optimize the use of energy by providing fine-grained control of air conditioning systems in underground malls and building blocks based on the number of people in public spaces, their location, and their movements. We can use the same information during a disaster and provide each person with the best evacuation route from underground malls.

What are the challenges involved in collecting, integrating, and storing data?

—Please tell us about techniques to collect, integrate, and store data needed for CPS and the challenges you face.

Takasu One of the challenges of CPS is the diverse forms of data being handled. In addition to sensor data representing the physical conditions of the real world, data also includes images and text information transmitted by people through social networking services like Twitter.

With CPS, information obtained from the physical world, such as sensor data, flows in without a pause. As a result, the conventional method of steadily storing data somewhere doesn't work. Even if the pieces of data are small, it's not easy to distribute the overall data for storage because of its prodigious size. Besides further improving data compression techniques, we must also develop systems for selecting and retaining data.

Another challenge is the difficulty of integrating data, because they have different representations. In the past, when businesses such as banks merged, and their systems needed to be integrated, "name aggregation" of data managed by the systems became a headache. Although we would expect that the systems being integrated can mutually handle the same kinds of data, because their representations differ, we need technologies to integrate heterogeneous information. We also run into this problem with sensor data. In CPS, this is called the data stream problem. It must be dealt with under limited computational resources.

For CPS-related data, first, data are integrated using the "time" and "place" of their generation as keys. For example, if we filter data like geographic information and GPS information from smartphones and automobiles, as well as information posted to SNS, and integrate the data along a time series, then we should be able to understand traffic conditions on a more detailed level because information obtained from multiple data complements each other.

To analyze massive data, a sound strategy is to extract data according to their representation of events – where, when, and what. By cross-comparing records extracted from data that contain the same kind of events in the past, we can estimate answers to questions like, "If an accident happens at location A, what will happen to road conditions around it?" For NII's research project, we are developing techniques to detect traffic incidents in real time. To accomplish this, our technology seeks to understand road conditions during normal times from automobiles' GPS data, and it looks for cars behaving differently from normal conditions. Currently, we are using only GPS data from vehicles. I want to enhance our technology with the ability to take in information from a variety of angles by collaborating with research groups that are analyzing TV broadcast news and SNS information.

Imai Another issue, one from a different perspective, is the need for discussions on how to protect privacy. We need social consensus on how far we can go in obtaining data on individuals' locations, movements, and SNS contents, and on the acceptable scope for storing and applying the data.

What are methods of data management and analysis, and the challenges you face?

—Please tell us about methods and challenges involved in managing and analyzing data.

Takasu There are two kinds of data for analysis. The first is "storage-based data,"*1 such as data of past records stored in a database. The second kind is "stream-based data," where the flow of data is never-ending*2.

Storage-based data are captured, searched, and analyzed using the techniques of conventional database management systems, data warehouses, and statistical analysis software. To carry out sophisticated analysis of massive data sets targeted by CPS, we need to further improve processing performance. Also, to extract latent information in the background of data, we need to refine machine learning algorithms. Our research is working to develop algorithms for analyzing CPS data using a statistical model called a latent topic model.

For stream-based data, analysis must be carried out under stringent time and computational resource restrictions. Because of this, researchers are developing CEP (complex event processing). This technique analyzes data in memory without the need to store it piece by piece in a database, and it can produce results in a short amount of time. Meanwhile, to analyze data streams, we need to conduct exploratory analysis while narrowing or broadening the scope of the target data. Right now, there is still a big technological gap between what is needed to analyze data streams and the technologies for doing so. Filling this gap will be a major challenge going forward.

Atsuhiro Takasu
Professor, Digital Content and Media Sciences Research Division, NII
Chair, Professor, Department of Informatics, School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

Imai It is important to set a clear purpose for CPS and work toward it. We need to clarify what data are necessary to meet the purpose, and think holistically to figure out conditions for managing the data, including data protection. At NII, we are working to solve issues like the ones we just described. We are working to create a foundation for CPS that will benefit society, public services, disaster response, and other areas.

(Written by Masahiro Doi)

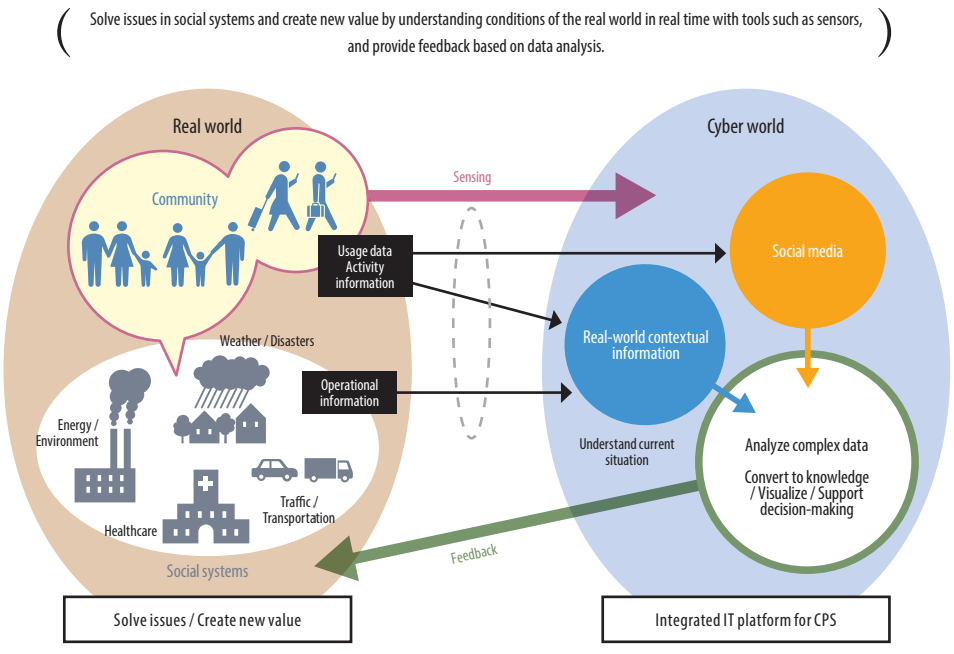


Figure: Social Cyber Physical Systems (CPS)

*1 Data used for purposes where the time difference between data generation and analysis is inconsequential, such as processing by general business systems, log management of IT devices, and search of SNS content. The data is stored first in a database, and then analyzed and processed.

*2 Data used for purposes where processing must be done in real time, for example to provide decision-making support and control of equipment based on rapidly changing conditions. The data to be processed comes in a flow that cannot be stored, and must be analyzed as soon as it is produced.

Attracting Visitors to Tourist Sites by Leveraging Mobile Phone and Smartphone Data

Understanding tourist movements to benefit regional development

The Japan Tourism Agency (JTA) is conducting research to discover new tourist routes and locales by analyzing big data from about 700,000 cell phone users. Obtained from the GPS (Global Positioning System) of their cell phones, the data aggregates information about their movements. Hiroyuki Kawataki, director at Japan Tourism Agency; Hiroaki Morinaga, director of Sasebo City Tourism Promotion Bureau; and Associate Professor Kenro Aihara of NII, chairman of JTA's research working group, discuss the purpose of this research effort and its methods, as well as future developments.

Information on people's movements gives birth to new tourist sites

Aihara Research of tourist movements being conducted by the Japan Tourism Agency (JTA) utilizes information compiled from the GPS data of about 700,000 cell phone users. This data is collected by mobile phone companies with the permission of the users and does not include their personal identifiable information, such as full name and home address, or even their basic attributes, such as age and sex. JTA analyzes data such as the area from which tourists depart, their travel routes, length of stay, and whether they lodged overnight. The results are provided to cities for their own applications. What purpose does this research serve?

Kawataki The number of domestic tourists has been declining in recent years. Foreign visitors, on the other hand, topped 10 million last year, the most ever. The tourism industry is facing massive changes, both quantitative and qualitative. There is a shift from group tourism to individual tourism, and needs of tourists are diversifying. At JTA, we are thinking of ways to invigorate the tourism industry by moving from excursion-based or day-trip visits to overnight stays and long-term exchanges. To create strategies for this shift, we need to conduct detailed surveys of tourists' needs and analyze the market. Instead of depending on our past experiences, we have begun research that utilizes objective data in new survey methodologies to understand and accurately analyze tourist movements.

We are conducting the new surveys in eight regions right now. We are targeting six tourism areas, as well as Fukushima Prefecture and the surroundings of Mount Fuji. The six areas are expected to attract many tourists by partnering with a number of towns and cities like Furano City in Hokkaido and Sasebo City in Nagasaki.

Aihara When you analyze data, you can even understand the movements of individual travelers. You can discover unexpectedly popular spots. If new tourist routes can be identified, you can attract more tourists to a location by adding the routes to visitor maps and other materials.

Mr. Morinaga, as someone responsible for promoting tourism in a city, what is your view of this research?

Morinaga To learn travelers' routes and their level of satisfaction, we had been interviewing visitors in 14 locations in Sasebo City, with the goal of collecting 10,000 responses. Not only was this method costly and time-consuming, but analyzing the data was also difficult. Our new surveys replace human labor with GPS data. We have extremely high expectations for their results. They can inexpensively and easily assimilate information, even information we have not discerned before.

Understanding actual conditions with data is effective for marketing

Aihara What kind of data are you focusing on?

Morinaga In Sasebo, a major challenge is figuring out how to get visitors to travel between our two major tourist spots, Huis Ten Bosch and the Kujuku Islands. To devise strategies, it is important to think about tourists' movements, such as where they come from, where they are going, where they stop to eat lunch, and so on. Their direction of travel is also related to physical sites posting tourist information. I think we can create

Hiroyuki Kawataki
Director, Regional Development Division
Japan Tourism Agency, Ministry of Land, Infrastructure, Transport and Tourism

Hiroaki Morinaga
Director, Sasebo City Tourism Promotion Bureau

Figure: Flow of Japan Tourism Agency's tourist behavior research utilizing big data

effective PR campaigns and business opportunities by clearly identifying the areas being targeted by tourists.

Aihara There are opinions about the impact on surrounding areas when tourists concentrate in a popular location because of influences like the Internet.

Morinaga There are two theories about what happens when an area becomes a big tourist attraction. One is that visitors will flock there and stop going to surrounding areas. The other theory is that a major tourist spot will draw new visitors and benefit its surroundings. If you can analyze the data, then you can know what actually happens. If the surrounding areas lose guests, then you can come up with a game plan more easily when you grasp the trend immediately.

I also want to know the tendency of visitors to stay consecutive nights in Sasebo City. I want to analyze their inclination and provide information on highly rated restaurants and appealing routes so that visitors will be induced to stay more nights.

Kawataki Every region is feeling the importance of being able to visualize its conditions. Even though towns and cities bear the responsibility for explaining what clear data shows, I think the visualization made possible by the data will also benefit marketing campaigns by the private sector and has significance for the entire tourism industry.

Morinaga Understanding routes taken by travelers will be helpful in creating new recommended courses, such as "the most popular walking course among 100 tourists." If we combine this course along with other courses in maps and visitor materials, then visitors will be able to choose from among the most popular courses "ranked" by 100 tourists. This is a springboard for thinking about new methods that will revitalize the tourism industry.

Data integration makes it possible to attach meaning to behavior

Aihara A variety of developments can be possible in the future by utilizing big data. For example, if we integrate GPS data with content from services like Twitter on a particular topic and tourists' attitudes toward it, we can attach meaning to their behavior. We can learn how they act based on their interests. And, by superimposing movement data from modes of transportation like cars, buses, and taxis, we can also understand these movements' conditions. Going forward, we can create support systems for tourism. Such a system can, for example, allow local businesses that welcome tourists to plan based on estimated demand in real time.

As you know, right now only data from tourists are being extracted. But there are many cases of businessmen and women touring for a brief period of time, during breaks on their business trips. Can we utilize information from such a group?

Morinaga Because groups that practice "sideline tourism" hold great promise, if we can extract data from them, we can utilize the data to study how to approach them and develop experiential programs that meet their needs. But the applications are not just limited to these examples. I think the big data we obtain will lead to the creation of programs that meet diverse needs.

Management that leverages IT is a major step for the tourism industry. But even if we have data, it is meaningless if we don't capitalize on it. How we use data to create tourist locales will be critical.

Cyber Physical System

Kenro Aihara
Associate Professor, Digital Content and Media Sciences Research Division, NII
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

Also, the data being gathered right now does not include personal attributes. However, age and sex are basic information for marketing. I would really like to investigate questions such as, "Is it true that women in their 40s travel often?" In the future, I think that if advanced filtering functions can be implemented so that attribute information can be obtained without compromising personal information, the quality of marketing can be further improved.

Kawataki Going forward, we will discover the latent needs of local communities by using collected data. At the same time we will extend our efforts to other regions and investigate other data that will be needed. To learn about the behavior of foreign tourists, I would like to gather information by asking them to download tourism-related apps.

Aihara Speaking of tourism apps, I was involved in the development of "Hime Pass,"* a travel support and service app being field tested around Matsuyama City in Ehime Prefecture. It does more than just present information. After you download the app, you can enter affiliated facilities as many times as you like for a fix price by purchasing excursion passes with the app. Besides users, the app also benefits local businesses.

Kawataki In any event, because GPS data analysis is not enough, we are studying not just the data itself, but methods to integrate it with different data. I want to leverage big data for the promotion of tourism by exploring concretely in what areas of tourism we can effectively apply these technologies.

(Written by Yuko Sakurai)

* Hime Pass
Smartphone app that allows users to enjoy a wide range of tourist sites in the greater Matsuyama area in Ehime Prefecture. After users download the app to their smartphones, it can present information that meets their needs simply by reading "markers" (NFC tags and QR codes) found in about 20 shops in commercial districts. URL: http://himepass.jp/ (As of March 2014, passes to tourist attractions are not being sold on the app.)

Creating Rules to Make Full Use of Personal Data

Balancing privacy protection and usefulness of data

The spread of CPS will lead to the collection of massive amounts of environmental data. However, the data may identify people and expose their personal data. Underlying the advancement of this technology is the creation of rules that seek to balance both the utilization of personal data and the protection of privacy. The development of these rules is now moving at a faster pace. Professor Ichiro Satoh of NII, the chair of the technical review working group studying technologies for the utilization and protection of personal data as part of the revision of Japan's Personal Information Protection Act, shares his views on the utilization of personal data in the future and the protection of privacy.

Ichiro Satoh
Professor, Information Systems Architecture Research Division, NII
Professor, Department of Informatics, School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

Utilization of personal data and protection of privacy are inseparable

With the rise of CPS (Cyber Physical Systems), a closer connection is being formed between the information world and the real world we live in. We are already seeing signs of this trend.

Take, for example, car navigation and map services using GPS (Global Positioning System). These systems capture the positions of automobiles and people. The information obtained can be utilized not only for users, but also for a wide range of applications, including improving the efficiency of social infrastructure and developing marketing strategies. Using CPS to gather and analyze personal data is beneficial for not just public agencies, but also the private sector. For example, if train passengers' daily activities can be understood in detail from their ride histories, then not only can this information help railway companies operate more smoothly, but it can also support shops near train stations in their creation of marketing plans.

However, the real world being treated by CPS includes people, and the information gathered by these systems can reveal some personal information about the users. When surveillance cameras installed on street corners capture and record images of people walking by, including their faces, data that leads to distinguishing and identifying individuals is being accumulated. Diagnostic records at medical facilities can lead to learning major secrets of individuals. To reduce the ability of data such as purchase histories to distinguish customers, we need measures that hide not just their names, but also information identifying individual products, such as model numbers. However, the value of such data as information is reduced.

As we face these issues today, an effort is underway to revise the Act on the Protection of Personal Information ("Personal Information Protection Act"), enacted in 2003. The Japanese government plans to establish fundamental principles for the revision by June 2014, and amend the law in 2015.

Discussions on "anonymization" for utilization of personal data

Professor Ichiro Satoh of NII participated in the Cabinet Secretary's "Study Group on Personal Data," convened in September 2013 to create a blueprint for the revision of the Personal Information Protection Act. Furthermore, he chaired the study group's "Technical Review Working Group," and studied issues such as the anonymization of personal data. The results of this examination were released as the "Technical Review Working Group Report" in December 2013.

Professor Satoh says that the issue of privacy protection is a critical one for CPS. "Because CPS technologies are spreading, it is becoming easier to identify individuals from obtained data. This can lead to privacy issues. To address them, the Technical Review Working Group discussed possible measures based on the newest technologies."

Professor Satoh says that when CPS is fully adopted, sensors will have wider use than they do now. Businesses that gather and use data will also increase. Integration of sensor data with other sources of massive data, like the Internet, must also be taken into consideration. If adverse effects arise as a result of using the latest technologies like CPS, then they must be

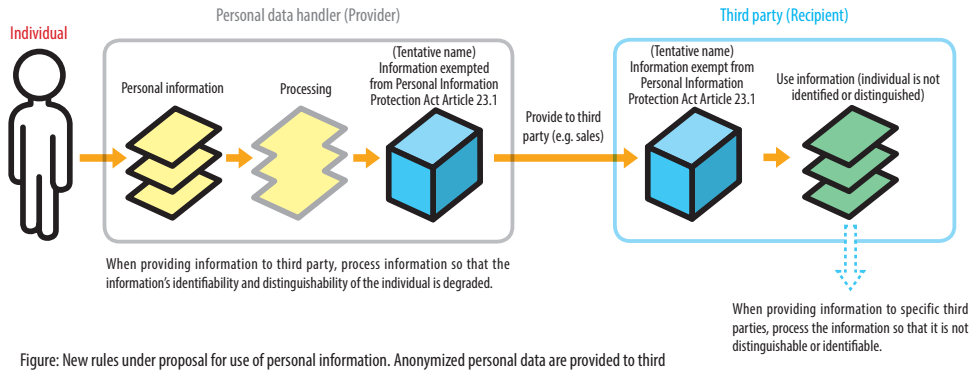


Figure: New rules under proposal for use of personal information. Anonymized personal data are provided to third parties as an exception to Personal Information Protection Act, Article 23.1.

resolved. Dealing with such issues is just as important for people involved in technological development as research related to CPS itself.

Because of these issues, the aforementioned Working Group Report examined clarifying the scope of personal data, anonymizing it, and providing data to third parties to protect privacy information from a technological standpoint.

As CPS and big data technologies advance, the problem of being able to identify individuals in data emerges, even for data that separately does not pose privacy risk problems. This can happen when information about the real world obtained from sensors is integrated with diverse information on the Internet. The Working Group Report therefore states that it is necessary to explicitly consider the scope of personal data not in terms of "identifiability" – that is, whether one can determine the identity of the individual to whom some data belongs – but, rather, in terms of "distinguishability" – that is, whether one can know that some information is related to an individual.

For example, even if a purchase history is anonymous, you may be able to identify the person who made a certain purchase if, for example, the item purchased was a minor product and you frequent his or her blog. You may also be able to determine other purchases he or she has made. In reality, it is not unusual for online merchants to use customers' names and search their Twitter and Facebook accounts to determine their address, interests, and behavior.

Anonymization is gaining attention as a method to preserve privacy from data when the data is published and analyzed by different parties. Anonymization is a set of technologies to reduce the identifiability and distinguishability of individuals by processing personal data. However, as Professor Satoh observes, "We can't place too much expectation on anonymization." The Working Group Report states, "It is not possible to define a reasonable level of universal anonymization that completely neutralizes the ability to distinguish (or identify) individuals from information."

Meanwhile, protecting personal data and utilizing them involve tradeoffs. This means that personal data that tamps down identifiability and distinguishability has reduced value for its utilization. In the case of purchase histories mentioned above, reducing distinguishability means the need not just to hide the names of customers, but also information identifying individual products, for example their model numbers. This results in the information losing value.

Case-by-case handling is necessary

If personal data is anonymized and the identifiability and distinguishability of individuals are reduced, the path is open to the utilization of personal data. Businesses utilizing personal data may have expected that the release and application of explicit, written guidelines on anonymization would advance the utilization of personal data and lead to the exchange and analysis of data between businesses. However, the Working Group Report that studied the issues surrounding privacy from a technological perspective did not seek to resolve them uniformly with guidelines. It pointed out the need to handle privacy issues on a case-by-case basis.

Furthermore, anonymization so that "each individual cannot be identified or distinguished" is not easy. The report discusses ride histories on transit IC cards like the Suica rail passes. Measures to reduce distinguishability include deleting the "ID" field that ties the data to individuals and removing data from stations with few embarkations and disembarkations of passengers. At first glance, these measures seem adequate. However, a ride history is composed of stations where a passenger embarks and disembarks. Even if a passenger moves from one heavily-used station to another, it is not necessarily the case that he or she will belong to a large group of passengers for this combination of stations. If combinations of some stations contain only one passenger, then those persons may be identified. If so, the measures described above cannot be said to be adequate. Furthermore, because the number of passengers changes daily, there is also the need to change the system of anonymization. "Unfortunately," stresses Professor Satoh, "not all businesses carry out appropriate anonymization. We need to design a system that takes this possibility into account."

Can personal data be effectively used?

Providing personal data to third parties is premised, of course, on removing "identifiability" and "distinguishability" from the data. However, if anonymization is difficult, providing data to third parties also becomes difficult, and the utilization of personal data cannot move forward. Thus, the Working Group Report defines a new type of personal data for provision to third parties. The proposal envisions a legal framework that allows personal data that is anonymized to a certain extent, such by removing personal names, to be provided to specific third parties as an exception to the Personal Information Protection Act. The businesses receiving the data are bound by law to not identify individuals associated with the information.

For the revision of the Personal Information Protection Act, a measure being discussed is the establishment of a third-party agency as the commissioner of personal data. The commissioner would monitor the uses of personal data to make sure they conform to the law, and take necessary action including audits if violations are discovered. Participants discussing the revision of the act also state the necessary of establishing a technical body to deal with anonymization technologies.

Professor Satoh says, "When thinking about 'offense' and 'defense' from a technological standpoint, 'defense' is frequently more difficult. Computer security is a classic example. The same can be said for privacy protection. If the demand for data scientists increases, so too will the demand for technologies supporting anonymization and engineers who implement them."

Case-by-case handling of personal data to deal with the difficulty of both their utilization and protection of privacy will be a major challenge for researchers and engineers going forward. In other words, personal data may become their new frontier.

(Written by Akio Hoshi)

At the Forefront of Compression/ Decompression Technologies

Essential technologies for smart use of sensing data

To efficiently store and carry out high-speed processing of massive data collected from sensing devices, compression/decompression technologies are essential. However, conventional methods present a variety of processing issues. To resolve these challenges, Associate Professor Kunihiko Sadakane of NII and Associate Professor Takuya Kida of Hokkaido University are conducting joint research to develop more efficient compression/decompression technologies.

Challenges of efficient compression/ decompression of sensing data

In recent years, sensing data obtained from devices such as GPS (Global Positioning System), cameras, mobile phones, and smartphones have drawn attention from big data researchers. To realize CPS (Cyber-Physical Systems), methods that can efficiently store massive data continuously collected from these myriad devices and freely extract them for use on computing platforms are essential. To develop these systems, more effective compression/decompression technologies are required.

Conventional compression/decompression technologies face limitations when it comes to efficiently processing and rapidly searching a large volume of data. Associate Professor Kunihiko Sadakane of NII explains: "The problem with current compression technologies when performing high-speed processing of massive data is that they can't read/write compressed data stored in computer memory as-is. In addition, while random access of specific data is easy with non-compressed data, it is impossible with compressed data. Each bit of compressed data must be decompressed from the beginning. To handle this problem and reduce decompression time, the data must be divided into multiple blocks, each of which must be compressed. This leads to a loss in compression efficiency."

Associate Professor Takuya Kida of Hokkaido University adds: "If the amount of data is small, it can be processed as-is in memory without undergoing compression and decompression. However, for high-speed processing of massive data like sensing data without compressing and decompressing the data, you need a large-scale computing environment equipped with a high-capacity RAM. What's more, to store data you naturally need to increase the hard disk space. In the past, methods were adopted to deal with these constraints, such as aggregating only data with values near a certain constant, storing only averages, and so on. However, these processes resulted in knowledge being overlooked, as data containing the knowledge were culled."

Associate Professor Sadakane says, "To solve such a problem, I have been conducting joint research since 2012 with Associate Professor Kida and Professor Masayuki Takeda of Kyushu University to research basic technologies that will enable more efficient compression and decompression."

Improving compression with VF coding and Re-Pair algorithm

To realize efficient compression/decompression, what joint research efforts are the professors engaged in? Associate Professor Kida is investigating compression/decompression using VF coding (Variable-length-to-Fixed-length codes) and the Re-Pair algorithm. A VF code is a coding scheme that assigns a fixed length codeword to substrings of different lengths in the original data.

Associate Professor Kida explains: "For example, if a block of data is expressed with 5 bits, and you want to extract only information expressed by the fifth codeword from the compressed data, you just need to extract the 21st to 25th bits. If the data were compressed by variable-length codewords, you must look in order from the beginning."

Takuya Kida
Associate Professor, Division of Computer Science
Graduate School of Information Science and Technology
Hokkaido University

Figure 1: Problems encountered with conventional data structures

The size of data is big	The size of the index is too big
● Random access is not possible (slow) when data are compressed. ● Index must be added to perform search	● The data cannot be stored in memory. ● The data must be stored on disk. → Slow processing ● Need to reduce size by reducing amount of information in index. → Reduction in precision

Succinct data structure

Conventional index (suffix array) + data	680.4G
Succinct data structure (index + data)	21.6G

Cyber Physical System

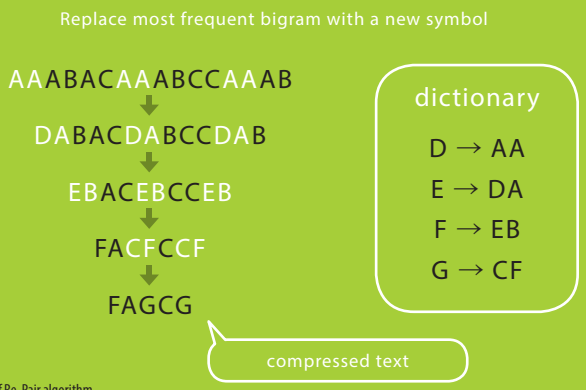


Figure 2: Scheme of Re-Pair algorithm

In this way, VF coding has the benefit of allowing easy access of compressed data because the boundaries of the codewords are clear. However, compared with variable-length coding, it is difficult to improve compression efficiency. Associate Professor Kida explains, "Because of this, we are working to create a system that can process functions such as high-speed keyword search without the need to restore data, all the while maintaining efficiency. Our approach is to combine VF coding with compression techniques based on the Re-Pair grammar conversion algorithm (Figure 2)." Combining these two schemes results in compression efficiency that exceeds the general compression tool "gzip" (Figure 3).

New Compression Techniques Using Succinct Data Structures

Meanwhile, Associate Professor Sadakane is advancing research on compression/decompression based on "succinct data structures."

Associate Professor Sadakane says, "I'm conducting research on a new compression scheme called succinct data structures. I've developed compressed suffix arrays that can allow fast in-place searches of compressed character strings." Associate Professor Sadakane elaborates: "When implementing a search index for character strings on a computer, extra data is often added. As a result, the index often becomes bigger than the original data. To address this issue, the concept of compressed suffix array has been proposed. However, because the text itself is needed for the search, there was the problem that the index size did not become smaller than the text. I therefore proposed adding modifications to the original index so that the text itself is not needed by the search algorithm using compressed suffix arrays. I also proposed using an algorithm that decompresses the entire text or a part of it from compressed suffix arrays. Besides compressing text and indexes, my ideas allow search of an arbitrary phrase and partial decompression of any section of a text."

Associate Professor Kida says: "The compression/decompression technologies utilizing succinct data structures being researched by Professor Sadakane can make data compact enough to be processed in memory. Data access becomes easy as a result. In other words, high-volume data can be processed and analyzed without the need to prepare a large-scale computing environment with high-capacity memory."

Associate Professors Sadakane and Kida are currently conducting joint research on compression/decompression technologies that target text data. A practical application being anticipated is the processing of DNA and genomic data. Going forward, they will also apply their technologies to sensing data, with work expected to begin first on the compression/decompression of data gathered from automobile position information systems.

If new compression/decompression technologies make high-speed processing of sensing data possible, we will be able to obtain a variety of knowledge that, in the past, has been buried in massive data and overlooked. Anticipation is building for the prospect of new discoveries.

(Written by Hideki Ito)

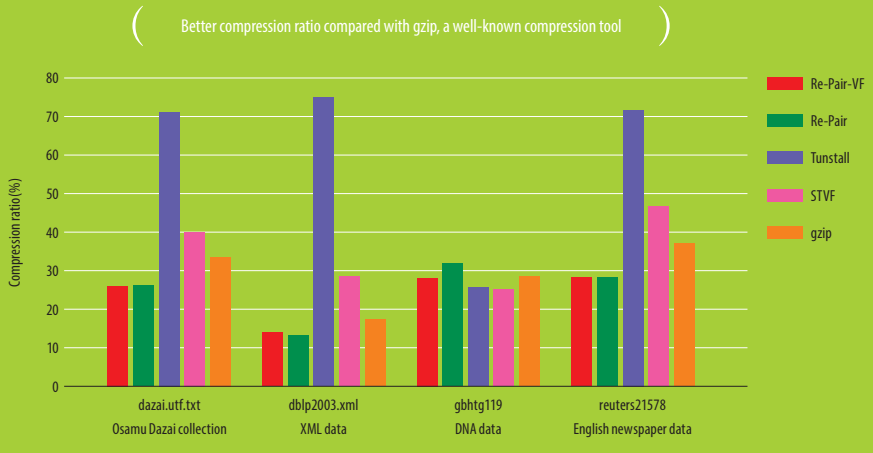


Figure 3: Improving compression efficiency by combining Re-Pair algorithm and VF coding

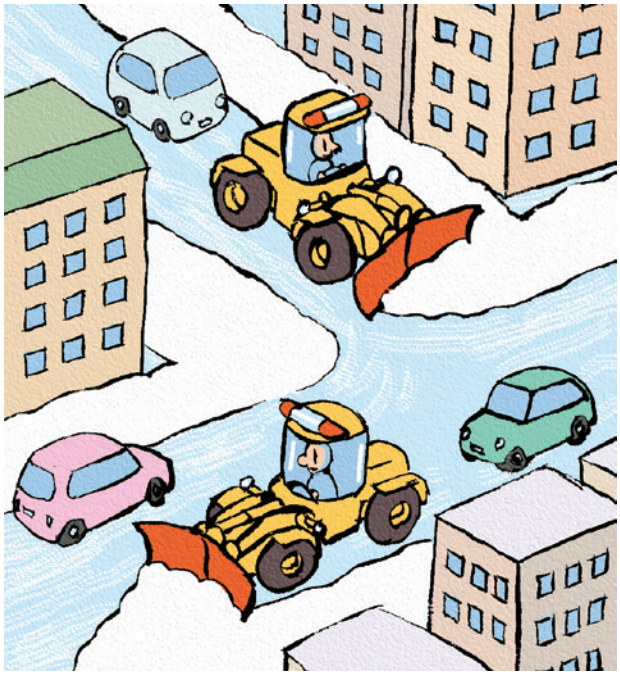
Kunihiko Sadakane
Associate Professor, Principles of Informatics Research Division, NII
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

NII Essay

The Operating System of Urban Society

Yuzuru Tanaka

Professor,
Graduate School of Information Science and Technology
Hokkaido University



Do you know which city in the world with more than one million people has the most annual snowfall? Sapporo. This city of 1.9 million people accumulates an average of six meters of snowfall per year. The distance covered by snow removal vehicles during one night totals the round trip distance between Sapporo and Ishigaki Island (Okinawa). Sapporo spends more than 15 billion yen per year for snow removal. Last year, this sum swelled to 22 billion yen. It is natural, then, to try to apply ICT to optimize snow removal routes and timing and to reduce cost. Actually, such an effort was attempted more than 20 years ago. However, it did not yield satisfactory results.

Meanwhile, ICT has undergone profound transformations during the last 20 years. Innovations include the proliferation of GPS-embedded smartphones and car navigation systems, the development of technologies utilizing location data, the widespread use of traffic sensors, and the advancement of technologies for distributing, storing, managing, searching, and analyzing massive data.

The physical world can be monitored by utilizing car navigation systems, smartphones, masses of sensors ubiquitous in cities, like traffic sensors, and weather radars. Also beneficial are databases of snow removal records. Snow fall and snow drift conditions can be measured by laser range scanners mounted on vehicles. From probe car data, we can estimate in real time the effects of weather change and snow removal histories on road icing and accessibility, and quantitatively evaluate which roads most urgently need snow removal. By analyzing the correlation with traffic accidents, we can also easily send up-to-the-minute warnings on dangerous road conditions to drivers.

Cyber-Physical Systems (CPS) seek to provide users with optimal control of the world they correspond with by modeling physical space in cyber space, coupled with the use of related databases. The physical world to which a CPS applies includes the human heart for pacemakers and large-scale plants for control systems. The topic of interest in this essay is a social CPS, which expands the world under the CPS' purview to the whole of urban society.

A similar concept could be found quite some time ago in Mirror World by the computer scientist David Gelernter, published in 1991. Mirror World refers to cyberspace that faithfully reflects the physical world – like a mirror. Gelernter proposed “tuple space” (shared memory space) as the interface that maintains consistency between the two worlds.

More than big data systems, social CPS is the operating system of urban society. It is its resource scheduler, and its virtual model to expedite the development and application of analysis and control systems. It provides a user environment that supports the agency of people in decision-making.

The need for social CPS in building sustainable, safe, and secure urban societies is growing. The prerequisite basic technologies are maturing rapidly. Remaining efforts include opening data silos maintained by the private sector and the government, and analyzing massive, complex data that cannot be completely described by a single monolithic model. Social CPS is filled with tantalizing challenges for research and development.

Weaving Information into Knowledge

NII

Cover Illustration

When the real world is connected in real time with cyber space, you can tell at a glance what is happening right now and where. Want to go to Mount Fuji with your dog? Find information instantly on facilities where you and your dog can play together, trail conditions, and other enjoyable spots. The reality of CPS is at hand.

NII Today **No.49, Apr.2014** [This English language edition NII Today corresponds to No. 63 of the Japanese edition]

Published by National Institute of Informatics, Research Organization of Information and Systems http://www.nii.ac.jp/
Address: National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Chief editor: Yoh'ichi Tohkura **Cover illustration:** Makoto Komori **Photography:** Seiya Kawamoto / Yusuke Sato **Copy Editor:** Madoka Tainaka **Production:** Kurdyla and Associates Co., Ltd.
Contact: Publicity Team, Planning Division, General Affairs Department Tel: +81-3-4212-2164 Fax: +81-3-4212-2150 E-mail: kouhou@nii.ac.jp