“Development of Data Archive and Its Impact on Asian Studies”

Time 13:00-16:00, June 25, 2017
Venue: 31 Classroom, Main Building, West Campus, Hitotsubashi University

Shigeto Sonoda

Konnichiwa. I would like to start Kashiyama Seminar “Development of Data Archive and its Impact on Asian Studies.” Please find the brochure of Kashiyama Scholarship Foundation and handouts for today’s presentations on your desk.

First of all, may I invite one of the Trustees of Kashiyama Scholarship Foundation, Ms. Yuko Hatano, for her opening remarks?

Yuko Hatano

Good afternoon. Thank you for the introduction. I am the Trustee and member of the Screening Committee, Hatano, of Kashiyama Scholarship Foundation. Thank you for this opportunity for my brief remarks. Professor Marukawa, the President, and Vice President Sonoda of Japan Association for Asian Studies, it is a great honor for Kashiyama Scholarship Foundation to be able to co-host this seminar. The weather is a bit rainy, but thank you for joining us. In the handout, please find the brochure of Kashiyama Scholarship Foundation.

That includes the pamphlet brochure of our foundation and there is a copy commemorating the 40th anniversary, plus last year, the 11th Junzo Kashiyama Award pamphlet, and for this year, the award related information and the application form are included. In 1977, we started this scholarship program. At this moment, we are in the 41st year. In commemoration of the 30th anniversary, Junzo Kashiyama Award was initiated, and this gives award to excellent publications on Asian studies, and this is the 12th award. Professor Akira Suehiro, who is in this venue, is also kindly member of the Screening Committee.

From April of last year until the end of June, out of the publications in Japan and books on Asia studies, they will be qualified. The deadline is the end of June. A large number of publications are already being nominated for academics and researchers like you here today, probably you would have knowledge about excellent publications that we are not aware of. So, we look forward to any nominations of excellent publications.

Shigeto Sonoda

Thank you very much. Then, I would like to make a brief introduction of why we try to organize today’s session.

It is not necessary to say that environment of our research on Asian studies has been changing, and even the scholars who are using qualitative data are getting a lot of benefits from the development of ICT.

For example, at the time of checking the literature, you can easily get it on website. Such technology has direct or sometimes indirect impact on the research agenda or methods that we are utilizing.

The second one is a more substantial thing; that is globalization of local issues. There are so many movements in Asian studies, for example, some scholars in China have come to argue so-called about the possibility of Global China studies, meaning that the China issues in many ways has become more of a global issue. Thus some issues are becoming simply international and global.

But today, I would like to focus on one aspect that was not previously discussed in the Annual Meeting of Japan Association for Asian Studies; that is the increasing number of data archive projects in Asia. Of course, there are many ways of doing research in Asian studies, but many orthodox area study experts still heavily rely on qualitative data, including observation, document analysis and so on.

Whether we like it or not, many governments or NGOs have spent a lot of money and efforts to create and promote the archival projects. Today, we invited some experts who joined such movements to create their own data archive so that not only the experts but also the academicians as well as the practitioners can utilize the data, whether it be the kind of second-hand data or the first-hand data, to do whatever the purposes they have.

As I would like to ask each speaker to speak about their own experience of their national archival project in the first round, I would like to briefly mention about the name of speakers and information of their affiliation. First speaker is Professor Kim, Seokho from Department of Sociology, Seoul National University. Next speaker is Professor Wang, Weidong from Department of Sociology, Renmin University of China. The third speaker is Professor Ronald Holmes from Department of Political Science, De La Salle University. He is also the President of Pulse Asia, which is a very big research institute in the Philippines. And the last speaker is Professor Yoshino, Ryozo from the Institute of Statistical Mathematics in Japan.

We also invited two discussants to pick up several issues that would not be touched by the four speakers. One is Professor Tanaka, Akihiko, who is now the President and Professor of the National Graduate Institute for Policy Studies (GRIPS). The other is Professor Yamamoto, Nobuto, Professor of media studies, working in Faculty of Law, Keio University. As you know, two discussants here are important members of our association, too.

Now, I'd like to invite Professor Kim to explain his national project. Please welcome Professor Kim.

Seokho Kim

I am Seokho Kim from Seoul National University. My organization is Korean Social Science Data Archive, we usually call it KOSSDA. In this presentation, I am going to talk about the main activities and history of the KOSSDA at Seoul National University.

We are mainly doing data collection, data accumulation, archiving, dissemination and education. KOSSDA is one of the Korea's representative data archives, leading in the collection, dissemination, and promotion of research materials through various academic events and methodology education programs.

When we found this organization, we visited several famous data archive organizations, such as GESIS in Germany, ICPSR in the United States and CESSDA in Norway. At that time, we tried to take advantage of their experience in the history of data archive.

KOSSDA started as a non-profit social science library in 1983. We added survey data archive in 2003 and then KOSSDA integrated digital data and literature archive in 2006. These days KOSSDA is providing some social science data to the users as well as giving some valuable literature for the scholars in social science. We also provide the integrated online service of data and literature on Korea. In 2015, office of KOSSDA was moved to Asia Center in Seoul National University.

KOSSDA collects survey data, statistical tables, interviews and narrative history data, documents, observation records, and other kinds of data produced by research institutes and individuals. It established digital database and provides the data for the use in research through the Archive's website. In other words, we try to integrate quantitative and qualitative data into a single system.

Among our main activities, the methodology workshop, including the class for the advanced statistics, is very popular in Korea. More than 1,000 students from other universities take part in that workshop annually.

In 2016, KOSSDA acquired 191 sets of survey data and developed a database with 72 of those sets. As of February 2017, KOSSDA has accumulated 2,251 sets of survey data and offers 211 databases of qualitative data. This picture shows briefly the data archiving workflow from identifying dataset, selecting dataset to disseminating data.

As you may know, in Korean society since 1990, there has been a tendency to collect social science data lab. We have more than several thousands of social science data produced every year but we don't accept old dataset because of data quality. In this way, we're trying to identify the high-quality dataset among social science datasets.

As to data acquisition, as I mentioned earlier, we collect both qualitative and quantitative data. We acquired data from individual scholars and organizations. KOSSDA is a kind of consortium of more than 100 governmental research institutions and more than 50 universities. On the other hand, we contact individual researchers either through collaboration with National Research Foundation. We also search journal articles and we analyze the journal article in order to identify the high-quality data for our organization.

The foundation of KOSSDA's governance lies in the consortium of depositors in institutions. The consortium institutions not only cooperate in the acquisition of data but also play a vital role in the operations of KOSSDA through their official representatives. Right now, we have more than 111 research institutions in the consortium. We also signed MoU with almost all the institutions in Korea which collect social science data.

We don't archive the data as we receive. As soon as we receive the data, we try to process the data. Data processing means cleaning the data. So, we look through the data and we try to check whether there exists some logical inconsistency within the data. If possible, we try to check the original questionnaire.

After data cleaning and error correction, we construct metadata. Metadata is an introduction of the data or summary of the data. This is most important task for us, and in order to protect confidential information, we try to eliminate all personal information from the data. In qualitative data, there are lot of personal information and even history about personal experience, so we try to remove such information from the data.

When we construct the metadata, we use DDI (the Data Documentation Initiative) provided by German partner. This is a kind of standardization of data introduction in the world of social survey. We use DDI 2.1 version these days.

The data on politics and public opinion and the labor and employment are most popular in social survey of Korea. The most popular data is Korean General Social Survey. By using Korean General Social Survey, more than 700 articles were published in prestigious journals within 10 years.

Our website provides user with English service and we open the English website for the foreign researchers, too. We also signed MoU with foreign institutions such as Harvard University, Stanford University, and University of Chicago libraries. We also provide members in those institutions with data free of charge since 2010, as far as I remember, and we translated 250 survey data into English. Actually, we selected 250 data in terms of popularity and data quality. Thus foreign users can get access to the data archive freely.

We have more than 14,000 individual members in Korea and in other countries, and we also have 150 institutional members. They pay membership fee annually, and after we moved our office to Seoul National University, we are looking for some funding so that we can make the data service free for the public. But it's really difficult to find out right funding. The data archive project is all about money, actually.

We also provide Nesstar. What is Nesstar? Nesstar is an online analysis program. On the Nesstar, you can analyze the data directly from simple frequency analysis even to logic regression analysis these days. Nesstar can also make you use OLS regression, logistic regression, chi-square test, correlation, and so on. It's free of charge. On the Nesstar, we uploaded as many as 250 data which is better in quality and popularity.

We started Korean Research Memory (KRM) project since 2005. Korean Research Foundation supported KRM in collaboration with KOSSDA and we collected more than several thousands of data on the website of KRM project. On Korean Research Memory website, there are a lot of qualitative and quantitative data and they also upload all digitalized articles and books. It's thus a huge project.

KOSSDA has organized a large-scale research team to complete the publication of 2016 annual report *Social Trends in Korea*. We launched this project with Korean Statistics in 2008 and this *Social Trends in Korea* is most popular governmental report among public officers and scholars. This report shows the changes and trends of the Korean society these days. We also created Key National Indicators in collaboration of Statistics Korea these days.

We also have educational and training program.

KOSSDA offers methodology training programs developed by benchmarking the summer program in Quantitative Methods of Social Research offered by the University of Michigan's Inter-University Consortium for Political and Social Science (ICPSR), while also taking characteristics of Korean circumstances into consideration. Therefore it's really popular and we try to control the quality of the lecture in class. More concretely, we limit the number of one classroom which is 35; in fact, there are lots of competitions for students to register that summer educational program.

Every year, KOSSDA operates four methodology training programs divided into winter and summer workshops providing concurrent classroom learning and practical training. Spring and fall short-term courses, on the other hand, are composed of theory lectures.

We also organized some data fairs for famous Korean social science data. In Korean society, there are more than 55 huge panel data and we are trying to make some kind of networks kind of consortium among panel data producers. We are also managing that network to facilitate the sharing of their experience for making the data quality better. We hosted two data fairs this year. We also have the research paper competitions.

Finally, we are trying to make international network for East Asian – actually Asian data archive. In the last year, four major data archives from Japan, China and Taiwan joined the meeting in Seoul. KOSSDA organized an international conference with representative archive of three countries, SSJDA of the University of Tokyo, CNSDA of Renmin University of China, and SRDA of Academia Sinica, Taiwan, to discuss the creation of association of Asian data archives. As a result of this meeting, four archives agreed on the foundation of Networks of Asian Social Science Data Archives, which is called NASSDA, to set the directions for its activities.

For more than 10 years, KOSSDA had focused on quantitative growth but in the future we will be pursuing the qualitative development for the data archive. KOSSDA was established by learning after some major data archive centers in Western society, but in terms of the size and the activities, it's by far smaller. We need to learn more from the data archives in advanced countries such as ICPSR in US, ESDS in UK, GESIS in Germany. Thank you for your attention.

Shigeto Sonoda
Thank you, Prof. Kim. Next speaker is Prof. Wang from China.

Weidong Wang
Twenty years ago, one of the main problems of China was shortage of data. But since 2000, because of the growing funding support to the academic, more and more dataset is available in mainland China and the Chinese scientific community has realized the importance and the necessity for data sharing. Most important thing is that Chinese government as well as scientific research foundations in governmental sectors also has come to recognize the importance of data sharing.

Based on these historical conditions, we have CNSDA (Chinese National Survey Data Archive). The CNSDA is the first, and by now, the only social science data archive in mainland China. CNSDA is sponsored by the Natural Science Foundation (NSF) of China, the number one national government foundation in mainland China. The funding started from 2012 and will end by this year, but actually this funding is provided also by the Renmin University of China. We have some one plus one parallel funding.

After the end of this round's financial support, the President of Renmin University declared that he will continue to provide funding to this data archive project. But I think the Natural Science Foundation will also continue to provide the funding to this data archive. As part of CNSDA project, we started to provide online data access service in 2014, just 3 years ago. Thus our history is not that long compared with KOSSDA in South Korea.

The mission of CNSDA is to acquire and deposit the raw data and related documents of all survey projects implemented in mainland China just as every published book in China will be stored in the National Library in Beijing. Therefore, I will do further collection and documentation of the survey data and provide open access and other related services to the academic community.

The Principal Investigator is Professor Yuan, Wei. He is the former Executive Vice President of Renmin University and now he is our Director. I am acting as his co-PI as well as Acting Director of CNSDA, which is based on the National Survey Research Center at Renmin University of China.

Our team has one Acting Director - that's me - and four DBA and programmers, two UI and VI designers, six data managers, and so many part-time assistants. I guess we have more than 20 assistants now. Most of them are students and master and doctoral students.

Based on our philosophy, all the software of CNSDA uses open source software. The database we use is MySQL, the script language we use is PHP, and the web server we use is APACHE. All the system is based on the open source Linux system.

The most important point to build the data archive is how to get the data. Actually, before in 2012 when we started the project to search the available data resource in mainland China by searching all published journal papers, books and the reports. We found 712 survey projects implemented in mainland China from 1980 to 2012.

Out of these data, 286 are owned by Chinese government, 197 is owned in domestic universities and research institutes, 127 is owned by commercial companies, and slightly more than 100 is owned by some overseas scholars and institutions.

To build the data archive with some initial seed dataset, National Survey Research Center of Renmin University agreed to release rights of its main survey projects, which include Chinese General Social Survey, China Education Panel Survey, China Religion Survey, Chinese Longitudinal Aging Social Survey, China Employer-employee Matched Data Survey, China Enterprise Human Resource Survey, and College Students Panel Survey. All these seven datasets were first released in the CNSDA and that’s the only place where you can download these datasets.

As to this more than 700 datasets collected by other institutes, we used different strategies to acquire the permission to store the data in our datasets. As to the dataset owned by Chinese government, we visited government offices one by one to introduce what we want to do and tried to persuade them to store the data in our dataset, but by now, almost no success but in just one or two cases we got permission. That’s a real story. And as to the domestic universities or research institutes, we sent email or visited their website to get phone number to call them. We got some positive feedback but not that much. And we are talking about the opening of the dataset to all.

We have transferred the server room to some clouding storage. A Chinese clouding company, Alibaba clouding company, agreed to provide free clouding service for our data archive. That’s very helpful for us.

By now, 122 survey datasets can be searched or downloaded, and most of the data and documents are stored in our dataset. But some dataset just have information of the link to our own dataset. After I started to act as Acting Director, I tried to simplify the procedure for the register process. Today if you want to use the data, just register online, and then you can download the data right away just within 1 minute.

In addition to this, we also provide online Q&A self-sustained knowledge system. That will be very helpful to our users and we provide email, telephone consulting service. We also provide some restricted data service, for example, the geocode information for the users. We can help them matching our data with some census data and some macro-enabled data. Most importantly, all services provided by our data archive are totally free.

As of now, the number of registered users of CNSDA is more than 30,000 and the number of download is more than 200,000 as of the end of last year. The top 5 downloaded datasets are Chinese General Social Survey, Chinese Private Entrepreneur Survey, China Education Panel Survey College Students Panel Survey and China Longitudinal Aging Social Survey. The former two datasets were provided by the Chinese government agency, and I was the PI of these survey projects.

Users of our data archive are from different disciplines. By now, the disciplines with more than 200 users are 11. Economics, management science, and sociology are the top 3 disciplines; statistics, education, journalism, political science, law, philosophy, psychology, ethnology / anthropology are the next top 8.

The data used by overseas users are more than 22%, and eight countries or districts that registered users have more than 100 users. Next to mainland China, the second is United States, where the number of users is more than 2,000. The next are the United Kingdom and Japan. Users from Japan are more than 500. I don't know who they are, but I can check our database in order to know it. Then come South Korea, Hong Kong, Taiwan, Germany, and so on.

Shigeto Sonoda
Thank you very much. Next speaker is Prof. Holmes.

Ronald Holmes
After listening to the presentation by Professors Kim and Wang, I am pretty much aware that they are far ahead. They are like in the first world, and we're like emerging economy with regard to data archiving in terms of the number of servers and facilities we can use in the Philippines.

I am envious about China and Korea because both of them are based in universities. In the Philippines, if you look at survey research, survey research has a long history. I don't know when survey organization started in Japan, but the survey organization started in early 1950s in the Philippines, which, perhaps, is the first Southeast Asian country to have survey organizations. But it is an interrupted history indeed. It surfaced again in 1985 when the Social Weather Stations (SWS), a private, non-profit, non-partisan organization, was established. Pulse Asia, our organization, is relatively younger. We were established in 1999.

Professor Wang talked about resource constraints on the part of the center, but our own constraint is that we are generating our own funds and we cannot rely on universities or government, especially when government doesn't like survey results. In fact, they don't like them all the more specifically with the president we have right now.

I have communicated with my counterparts in SWS. They have 566 datasets and that's divided between 276 national datasets, which means national surveys. The largest survey they conducted had a sample of 57,000 respondents, which was an exit poll done in 2016. The regular size of the survey of SWS, similar to Pulse Asia, is 1,200 nationally, and then they have 290 sub-national surveys.

Five of their national surveys are dedicated to comparative surveys, namely, World Values Surveys and Asian Barometer. They have some national surveys where comparative survey groups piggyback on that national survey what we referred to as rider questions. So there's a module in the survey that they conduct wherein the ISSP or the Comparative Study of Electoral Systems will bring in some of their questions. SWS conduct regular quarterly surveys and their data is deposited in Roper Center in Cornell University, ICPSR in University of Michigan. They are about to join another international data archive, which is called I-ASSIST.

Since we are younger organization, we only have 251 quantitative datasets. We have only about 13 of our own qualitative studies; thus most of our researches have been quantitative. 65 of them are national surveys and the rest are sub-national. We also conduct our regular quarterly surveys and our survey data is deposited with the Roper Center in Cornell University. In this point, our organization shows difference from Chinese and Korean cases. We still need to depend on our former colonial master, the United States, which help us in terms of archiving our data because we really don't have enough number of full-time staff to work on data archiving.

A regular question that we ask is something more stable; that is, quality of life. We ask respondents whether they think quality of their living has improved or worsen over the past years and we ask respondents whether they think the quality of their lives will be better or worsened in the following years. If you look at this dataset alone, you will note that Filipinos would always say that the quality of their life has worsened over the past years but it will be better in the following years. Filipinos are, in short, losers in their life but optimist in nature.

We have asked respondents whether they're in favor of political party, too. Interestingly no more than 10% answer they favor political party. Why Filipinos do not favor political parties? Because Americans created a political system where parties are not so important! Our parties are just – as an American political scientist points out - internally mobilized. On the other hand, although voting is voluntary, 80% of the Filipino informants participate in the election and they still believe that elections promote democracy.

The Philippines is a democracy and the questions we put in the questionnaires are not found in World Values Survey or in Asian Barometer. We sometimes have an open-ended question for them to define what is meant by democracy to them. The most important and the most common answer they say is “secure employment” or “better pay”, which has nothing to do with democracy but more or less an outcome of democracy.

Shigeto Sonoda
Thank you. Next speaker is Prof. Yoshino.

Ryozo Yoshino
I am working for the Institute of Statistical Mathematics. Our institute does not have data archive, but we have a long history of conducting social survey.

Our institute was established during World War II, so you could imagine what they were doing in the wartime. After the World War II, Japan was occupied by US military. Under that occupation, our institute got a new mission to reorganize governmental statistics and also scientific methodology to carry out public opinion research for the development of postwar democracy.

Principal Investigator in Renmin University of China, Professor Yuan, was a visiting professor in our institute in 1990s. In those days, we discussed the idea of data archive in Japan and actually Professor Yuan got that idea and going back to Beijing and he immediately opened his center and then carried out Chinese version of National Character Survey. President Park of Gallop Korea also visited our institute in 1970s and 80s. He studied survey methodology and went back to South Korea and then opened the Gallop Korea. Thus we have a sort of connection in many ways.

After World War II, our institute helped reorganization of governmental statistics and public opinion survey, and by using the methodology of public opinion survey, we started Japanese National Character Survey that lasted over 60 years. You might know General Social Survey in United States or Eurobarometer in Europe, but all those surveys were inspired by Japanese National Character Survey that started in 1953, so this is the original one affected all similar surveys all over the world.

In other countries they opened data to the public, so they become well-known to all over the world. But our institute has not yet opened our dataset to the public. Of course we are publishing many summary reports to distribute all over the world, but we cannot say they are well recognized.

Around 1970, we extended some survey to comparative one. I am working for this survey for some thirty years and here is the partial list of our survey. Within these ten to fifteen years, we are covering some Asian countries, too.

Let me give you one example of a repeatedly used question in our questionnaire. “If you were born again, would you like to be born as a boy or girl?” Over 60 years, Japanese men’s response has been stable; some 90% of Japanese male informants say “boy again”, while only 5% or 6% of them say “next time girl.” But on the other hand, here is women’s response. In 1953, 64% of Japanese women answered “boy next time” and 27% of them said “girl again”, but that’s steadily changing, and by about 1993, these figures are totally reversed.

Studying cross-national comparability is our task. Each country uses its own language and survey methodology. In Japan, we could use almost complete list of residents or voters, so by using them we can conduct ideal random sampling. But in other countries, even in the United States or European countries, they cannot do so due to the lack of voter’s lists. They made a big failure in the prediction of USA presidential election and Brexit voting in UK due to difficulty of random sampling.

I will not go into detail, but for cross-national comparability, we developed a scientific paradigm to compare nations, which is called the Cultural Link Analysis or Cultural Manifold Analysis. It wouldn’t be very meaningful to compare two totally different countries from the beginning. Therefore we start to compare two countries or two groups which share some aspects but have different aspects too.

About 20 years ago, EU was going very well. When we talked about idea of unification of Asian countries at that tijme, some people criticized that Asia is too diverse to think about unification. But now we are looking at lots of problems in EU and we may find some much better way to look at soft unification among diverse countries in Asia.

In the past longitudinal survey, we identified some fundamental dimensions of Japanese. Let me emphasize just one thing here. Interpersonal attitudes have been stable at least over the last half century and probably much longer. And most likely, the basic aspects of interpersonal attitude may be stable in any country over years. You could find much more interesting findings of our survey on our website.

We have recognized the difficulties on cross-national comparability of survey responses under the differences of cultures, histories, religions, social institutions on economy and politics, as well as differences of languages and survey methodologies. People’s response is a mixture of his or her opinion with his or her general response tendencies unique to gender, race, personality, and so on. Japanese tend to avoid polar answers and prefer middle response categories or “don’t know” answers, while French tend to choose critical or negative categories and Indians tend to choose positive categories, and so on.

Survey methodology of each country shows its own economic historical, political, social conditions. Before conducting survey data analysis, we should pay more attention to the differences of survey methodologies. At the time of data analysis, single indicator or single index is not reliable, so we should make use of multidimensional data analysis, putting together data from many question items in many countries together to look over topological relationship.

I believe our survey data can contribute to enrich Asian studies for our mutual understandings of the world or the world peace and world prosperity. But in order to promote "evidence-based Asian studies," we need to establish a system to continue longitudinal and cross-national surveys over the decades. But data collection and data archiving is not enough as of now.

We need to pay more attention to the cross-national comparability because of significant differences of survey conditions, including sampling methods, languages, institutions of economy and politics. A study of Asian way of soft unification may give a new perspective for policymaking. I hope some political scientists will approach such issues.

Shigeto Sonoda
Thank you Prof. Yoshino.

Now, I would like to put simple questions to all four speakers before going into the comment session. One question is, what do you think are most challenging facts or trends that you are facing? Some of you happened to mention about financial issues, but according to my own experience, one of the biggest challenges that we may face is the good utilization of the data.

Professor Tanaka as well as Professor Inoguchi of the Institute for Advanced Studies on Asia at the University of Tokyo promoted AsiaBarometer Project from 2003 to 2008. But one of the biggest challenges, in my eyes, was that even though there was a data but it was quite difficult to find users who might get benefits by using the data. Even though those who produced, compiled, and created the data are keenly aware of the importance, but it's quite difficult to attract promising young scholars who will use the data to write new type of papers based on their data analysis.

The second question, which has to do with first question, is where can you find some breakthroughs to overcome such hardship? Some of the speakers mentioned about the importance of dissemination or international meetings, but I wonder whether you have thought of any other ways to overcome the hardship you will mention.

Seokho Kim
As I mentioned earlier, the most difficult thing is funding, but fortunately our institution has been supported by some cultural organizations owned by some conglomerate company. We get about half a million US dollars a year, but in the course of managing our institution, we need another half million US millions dollars. So, we are looking for more funding in collaboration with governmental institution and some company.

Regarding the users, data archive is supposed to disseminate the high-quality data and help them publish high-quality article. In my presentation, I introduced some educational program and some networks for data fair. But the key, or I can call it the challenge, is whether we can let them use our data in exercising and learning advanced statistics. We are also running some programs for paper competition, which I think is also one of the effective efforts to increase the number of data users for our institution.

The third challenge we have is the cooperation among the related institutions.

Regarding Korean General Social Survey (KGSS), for example, the collaboration with KGSS is very effective, but the reason is because Professor Seok, Hyunho in Korea and Tom Smith at the University of Chicago are our partners. In fact, Tom Smith was my teacher and we launched the Korean General Social Survey together. As soon as the Korean General Social Survey was established, we launched the Korean Social Science Data Archive in 2006.

We have to get some valuable and high-quality data from other governmental institutions and scholars in universities. The problem is that their interests are all different. So, we have to find out some ways to make them cooperate with our institution. That's why these days we are trying to make some networks which make the institution collecting the panel data to share their experience and their know-how in producing their data.

Regarding the data searching, we also have the problem that each institution in South Korea wants to make their own data archive. Wherever I visit the governmental institution, they tell me that they're going to make their own data archive; but the real problem is that such attitudes will bring about smaller number of users for each data archive.

Our strategy is to help them make their own data archive. Korea Telecommunication provides us with free and large cloud storage for the server, and through that server we can share some inter-space of the website. We also try to help them or give them some techniques or experience accumulated in KOSSDA for these ten years.

Let me finish my answer by referring to the example of US General Social Survey.

US General Social Survey is collected by National Opinion Research Center (NORC), at the University of Chicago, and the users of that data are more than 10,000 a year. Moreover, more than 500 journal articles are published on the prestigious journal. Their strategy is to have their own website from which the users can download the data they like to get. What is more interesting is that there are five more websites which can provide the USGSS to the potential users in the academia. ICPSR at the University of Michigan, Roper Center at the Cornell University, GESIS in Germany, and SRDA in India also have some right to disseminate the US General Social Survey.

Thus my belief is that as soon as the data, especially large dataset, is collected, it should be owned by the public, by the society.

Shigeto Sonoda

Korean case suggests that many institutions are competing to take a lead, but it's a really paradoxical situation. That's one of the concerns that Professor Kim is suggesting, too.

Weidong Wang
China has been in a similar situation to South Korea.

We have so many academic survey agencies which want to be the number one. But in my understanding, it's not the biggest challenge. For China, the biggest problem is that it's very difficult to persuade the Chinese government to open their data. Actually, just as I reported earlier, Chinese government is the main data holder in mainland China. According to my own statistics, Chinese government owns more than 40% of data contents, or the survey data. It is only for all the micro-enabled data or census data that's only owned by the Chinese government.

Actually, I think the current Chinese government realized the importance to data open. Our current Prime Minister Li Keqiang said so many times to force the government agencies to open the data, but they don't follow his order, which is quite strange. Though our Prime Minister told so many times to order the government agencies to open the data, Chinese government officials still lack in motivation to open the data. They don't think they can get anything to open the data.

Another challenge is the competition among different disciplines, rather than different agencies. "Economic imperialism" must be a universal phenomenon in all countries. Economists emphasize the causal model too much, and they like panel data or longitudinal data much better than cross-sectional data. In fact, Natural Science Foundation of China and Social Science Foundation of China are dominated by the economists. Therefore, in recent years, review committees in these foundations have come to realize that they should provide more funds for panel data survey rather than cross-sectional data survey.

That's a strange idea, but anyway, Renmin University of China will keep on conducting sociological survey projects. So since 2010, there are so many academic survey projects in mainland China including Chinese Family Dynamic Panel Study in Peking University, Child's Health and The Aging Study and Chinese Education Panel Survey in Renmin University of China. Chinese General Social Survey may be the only cross-sectional, large-scale social survey in China, which provides us with a wide variety of variables on different social issues. Fortunately, last week one of our alumni donated huge amount money to our university, so we can continue Chinese General Social Survey rather smoothly.

Shigeto Sonoda
Thank you very much. Explanation about each country's hardship might be suggesting the uniqueness of different countries. Professor Wang happened to mention about "economic imperialism" in China, so I want to tell my own episode. When I was struggling with organizing this seminar, I tried to find research projects conducted by different disciplines in Asia. But I found it very difficult to find economists in Asia who are trying to

combine Asian scholars by creating common dataset. But sociologists and political scientists are rather enthusiastically doing these things. Professor Holmes, you are a political scientist, but do you have any comments?

Ronald Holmes

Social Weather Stations was started by an economist and they have done self-assessed poverty questions across the period of time. Luckily, we don't have "economic imperialism" which Professor Wang mentioned.

The first challenge in the Philippines is that we do not get much resource and warm bodies working on social surveys on a combination of qualitative and quantitative studies, because with those two you can actually be able to dig deeper into certain social concerns or phenomena. Unfortunately, we don't have that many warm bodies in the Philippines who can work with us and who can work on the data.

Every time we frame our survey questionnaire which takes us the longest amount of time, I searched through items of surveys done in other countries. But it is rare for me to find a survey item in China or Japan or South Korea because the survey questionnaires are written in their own languages. So more often than not I end up looking at pure researches, I end up looking at Michigan. So if NASSDA in South Korea can help us by translating some of these questionnaires into English, which allows access among at least in Southeast Asian countries, that would be great.

Shigeto Sonoda

As to the Institute of Statistical Mathematics, I think it is quite unique in the sense of its having a long history but it does have archive center. While this institute has been carefully promoting international comparisons, I'd like to hear the hardship or challenges that the Institute or Professor as individual is facing.

Ryozo Yoshino

As I said, we haven't opened our raw data to general public yet. But in the past 20 or maybe some 30 years, we are working to construct our data archive, not necessarily in our institute but in big institutes like GESIS in University of Cologne or in Roper Center in Cornell University. The size of our survey team was much larger but now we have only two or three members in our institute who will join social survey project, although total member of the institute is some 45. We are still carrying out big survey projects but it's really hard to do so because now is time of big data; engineers have no problem to get huge amount of money for their research.

But in our study, before World War II, in United States, prediction of presidential election outcome based on big data turned out to be a failure. So, we started appreciation for small sized but high-quality statistical data for prediction. But nevertheless, these days our engineers talk about big data, big data, and big data. Our Japanese National Character Survey lasted over 60 years but every time we start to conduct survey, we cannot ask money from government just for conducting longitudinal survey.

Therefore we pretended as if we start completely new project to get financial support.

As a researcher, you may understand or appreciate the importance of longitudinal survey for time series comparison. But talking about the National Character Survey, bureaucrat may say, “Just conducting National Character Survey is no good.” That’s the typical attitude of the bureaucrats. Therefore I am having hard time to get the money to carry out survey.

Nowadays, Japanese younger bureaucrats couldn’t understand why we are carrying out our survey. Recently we talk a lot about evidence-based policy-making all over the world. Once I talked about that problem with a young bureaucrat, he mentioned pessimistic views to support our surveys. Many researchers from US, European countries or other countries, on the other hand, come to Japan without knowing Japanese language, who ask us to pick up some statistical report to make use of it. That’s the natural attitude of our political scientists all over the world. But I wonder why Japanese bureaucrats are so different from them.

Shigeto Sonoda
Thank you very much. So some issues have to do with finance but the other issues have to do with support, including financial support by the governmental sector. Different countries have different issues, but again, we have to look at this project as an attempt to create public space where we can share the data for our researches.

Thank you very much for your responses to my tricky questions. Next, I would like to ask Professor Tanaka to make comments on the previous argument.

Akihiko Tanaka
Thank you Professor Sonoda and all the four panelists for providing us with great amounts of information which we need to digest. I would like to skip some of the propaganda I prepared and start with the topic I think relevant to today’s discussion.

Today’s presentations are mostly about survey data and survey data archives. But I would like to start with the statement that we are now confronted with abundance of data in various forms.

Economists are now able to use the entire set of World Development Indicators (WDI) with a single click of downloading it. WDI indicators are 1,504 indicators for 264 countries from 1960 to 2016 in a single CSV file of 189 megabytes, and similar data is available from various international organizations. IMF has provided us with International Finance Statistics and Direction of Trade Statistics of 3 gigabytes, which is also a huge CSV file. There are lots of national economics and statistics, too.

When it comes to economics, Hitotsubashi University where we are now has been long known for collecting long-term economic statistics in developing countries. Other than this, lots of survey data are now archived; World Value Surveys and various, various barometers around the world as well as national and media surveys have come to be archived.

In addition to these numerical data, we are now having lots of text data. If you visit newspaper companies' website, you will have entire dataset of their newspaper items from 19th century or nearly entire 20th century. Our humble database called "The World and Japan" also provides more than 8,000 documents, which are useful for international studies. If you visit Japan's National Diet's website, you can search the transcripts of Diet deliberation from 1945 up to now, including plenary sessions as well as committee sessions. There are lots of image data useful for social, economic and historical analyses. The Japan Center for Asian Historical Records is now compiling many historically important records from pre-war Japan with all the images of the original documents. Therefore I think we are now living in the age of abundant data.

The challenge for area specialists as well as social scientists is how to use them. I think that's the motivation that Professor Sonoda has guided us to assemble here. Paradoxically, however, there is a problem of so-called absence of data. At least among younger people, there is a tendency that if the data doesn't exist on the Internet, it doesn't exist at all. In other words, if you do not upload your documents or data numbers on the Internet, you will be neglected by the entire world.

I think there is some relevance to the plight of Japanese National Character Survey that Prof. Yoshino mentioned. I would really like to encourage Institute for Statistical Mathematics to put next round Japanese National Character Survey data open to the public with raw data. If you do not put them on the Internet, nobody is going to take care of it.

Similarly, there is a huge number of data that has not been fully used simply because they are not open to the public like limitless number of surveys conducted by the Japanese news media companies. Each newspaper company conducts a survey almost every month or every week, but they do not allow the public to use the data. Raw data are all hidden, and only limited number can use the date of, for example, approval rate of Mr. Abe. In other words, you have no access to the data to get to know what types of people are supporting him for what reasons.

Data can be absent because of the disruption of funding. For example, as to the AsiaBarometer project that Inoguchi sensei, I, and Sonoda-san had worked very hard, the survey ended its data collection in 2008 and after that the data collection has not been done at all. Even if you have valuable survey data, these may be terminated and maybe forced to disappear from the Internet. In such a case, it substantially means that these data cease to exist.

The next issue is on what I call tidy or messy data issue.

Normally, for the text data and image data, the most important use and purpose is to find out what you want to tell. If you could find out a document that supposes you in your monograph to prove that somebody was the villain, it would be sufficient. But, once you start to use the large data, you may come across with the issue of messy data and tidy data. As is explained in *Anna Karenina* written by Tolstoy, “happy families are all alike, every unhappy family is unhappy in its own way.” Later, Wickham and Grolemund changed this message into “tidy datasets are all alike, but every messy dataset is messy in its own way” in their book *R for Data Science* (2017). I think one of the problems that we can persuade the students and others to get into the business of analyzing large dataset comes from its messiness.

Usually, reading through the code book is a torture. As long as you are working on a single survey data, you may be able to accustom to the code book and you can use the dataset eventually. But once you go beyond the single survey to analyze the data with socioeconomic information provided by World Bank or somewhere else, then you have to deal with two messy data with totally different categories. Sometimes, even the names of the country may differ from one dataset to the other!

World Development Indicators use ISO code for different countries, for example, but their use in certain countries is different from IMF’s. Though the office for World Bank is located next to IMF, they are creating a small difference, which accumulates, accumulates, and then it becomes almost impossible to use both data at the same time. As Wickham said, 80% of data analysis is spent on the process of cleaning and preparing the data. Once you could successfully clean up your tidy data, your analysis can be done with 20% of the time left with powerful statistical programs, but this going through 80% oftentimes creates the challenge.

What are we going to do? We don’t know, but I would like to submit some of the things that we may do. It’s not particularly unique or particularly imaginative, but I think survey data should have consistent format and statistical data should have consistent format, too.

I think all over the workers of handling data should teach their students to improve their capability to analyze data first in their own respective disciplines. Probably we should start with a survey data. But later, we may go beyond the use of single survey data, combining survey data with socioeconomic data provided by various organizations. When we worked on AsiaBarometer project, many authors were asked to analyze AsiaBarometer’s data to write papers, but I’m afraid to say that it was a rather challenging task for them to combine AsiaBarometer data with data provided by other sources.

In many cases, survey data must be better utilized in conjunction with other socioeconomic data. Suppose you ask Japanese whether they are happy or not. Their answers might be “Well, I am not so happy.” But if you ask the same question to Thai, they might answer “I am happy.” Their response patterns might differ from those of Japanese, and I think the investigation into what causes such differences may suggest importance of some cultural variables or socioeconomic factors. These factors may not be obtainable from the survey data itself, meaning that you should go after the data provided by development agencies or governments.

The fourth problem is what we are going to do with this text database. Most of the text database is used for a single specific purpose of finding the right text. With hundreds of Prime Minister's speeches, hundreds of Foreign Minister's speeches and others, why not use those texts as a source of really important data to tell some big social and economic trend? Quantification of text, I think, is a challenge for the future development. Recently more and more computer programs can make a morphological analysis of languages like Japanese or Chinese. Previously, it was hard to analyze Japanese or Chinese because it's hard to tell which constitute word and how to separate sentences. But now morphological program is becoming much more capable. So we are able to make quantitative analysis of texts.

To sum up, creating tidy data out of huge chaotic assembly of big data is our challenge. Thank you for your listening.

Shigeto Sonoda
Thank you, Prof. Tanaka. Next, I would like to invite Professor Yamamoto for other comments to our argument.

Nobuto Yamamoto
I am wondering why I was asked to give a comment here because I am a quality-oriented researcher, not quantity-oriented researcher. I see myself as area specialist, especially Indonesian expert. I am not familiar with all these survey data analyses or big data analyses, although several of my students are now interested in doing such new type of studies. Thus I am here today to give you basic questions from the point of area specialist.

Although Professor Yoshino mentioned that nowadays bureaucrats are neglecting survey data, survey data is still useful for public policy as well as academics. Thus it's important to improve methodology, especially explanatory power, as Professor Holmes suggested in his discussion.

In the context of Southeast Asian studies, survey data research developed rapidly in the 21st century. It has to do with the political or sociopolitical conditions. It has started since mid-1980s, because of the democratization of the Philippines and then other countries like South Korea, Taiwan, and Indonesia followed. Under the changing political social conditions, survey data analysis has become important and popular, which was accelerated by the foreign government and foreign aid.

One of the major fields of studies of survey data analysis is election studies. In case of Indonesia, thanks to Japanese and the United States' official aid, they established survey institutes and started to conduct opinion poll survey. There used to be one or two survey institutions, but now dozens of survey institutions are there. They compete for business and they work for the clients. Sometimes they even manipulate data in the favor of their clients. Some political parties own their own data analysis survey institutions and they always provide a very important data for themselves, which might be useless for the public or the researchers.

This new tendency is now pushing us, us means the area specialists, to collaborate with the scientists or the manufacturers or the architects of the data survey. But it's hard to find such new generation of scholars. In Indonesia, those who conduct researches are educated in the United States or in UK in the 1990s or the early 21st century, who came back to Indonesia and started their own business because of the democratization. Unfortunately, however, they looked for money. They don't have any academic positions. They cannot find any academic positions in Indonesia as well as in other foreign countries, so they started profit making. It's sad in a sense, although they are very good well-skilled scholars.

It's hard to conduct cross-national survey, because most of data collection and analysis has been conducted nationally. They only work for the national purposes, and it's very hard to conduct cross-national survey, except some cases like Asian Barometer and Eurobarometer. You also need to have someone who specializes in one country or who knows about the country to analyze the data; otherwise it's really difficult to compare countries.

The third point is that survey research has become a big business in Indonesia for political purposes as well as business purposes. Some of the researchers I know always talk about methodology but their language is totally different from mine. Most of the time, I don't understand what kind of language they are talking about. But the question of epistemology is left out; how people perceive some topic or environment in such a way. The question that area specialists have been traditionally asking is how we can understand a country that we are studying. This is how the traditional area specialists approach to a research question. But this is something the survey data analysts must have overlooked over the years.

The questions for presenters are as follows. The first question is how to conduct cross-national survey. Second question is the credibility of the survey.

In the future, I think open access to the data will be very important especially for those who lived in academia in underdeveloped countries where infrastructure is poor. In case of Philippines as well as Indonesia, youngsters do not have free access to the Internet. They don't have advanced skills to use new models, either. They cannot conduct any research without outfit, so if you could make the data easy access to those youngsters and the scholars in the non-well-developed countries, that will open up new field of studies.

If you can create online forum of how to make a database or how to analyze database, that will create a new field of study, too. This is what I've already started in the case of social media like Facebook. People are creating forum to talk about political, social, economic issues. Infrastructure is already there. All we have to do is to provide data so that they can discuss with the new ideas. I guess this suggests the future direction of Asian studies. Thank you.

Shigeto Sonoda
Thank you Prof. Yamamoto. Now, I would like to ask four speakers to make some responses to the comments that two professors of our association have put.

Professor Tanaka mentioned importance of triangulation method, even though it’s quite difficult. Professor Yamamoto mentioned about the importance of international collaborations. I would like you to respond to the comments made by two professors.

Ryozo Yoshino
Thank you very much, Professor Tanaka. You encouraged us to open our data, and I will do that certainly. In case of what you call triangular strategy or…

Akihiko Tanaka
Triangulation method.

Ryozo Yoshino
By combing several different datasets, we can start new kind of analysis, -that’s true. That’s important in order to persuade people to construct data archive in Japan, which I emphasized several times. But in reality, with some social big data like those obtained in Google networks, we have to be very careful because degree of personal identification becomes stronger. Combining certain kind of data with ours may create another new problem, so we have to be very careful about it.

Shigeto Sonoda
Thank you very much for pointing out some negative aspects of combining different dataset.

Seokho Kim
Regarding the combination of different types of data, I would like to emphasize bright aspects of that process. As a social scientist, we are always trying to explain some social phenomena with the data. So, as long as we can protect the personal information, we can use hierarchical linear modeling in explaining the individual dependent variable based on the structural characteristic and personal properties and their attributes, for example.

KOSSDA is trying to provide some information about the local government. In Korea, there are 234 local governments and we are trying to provide users with the characteristics or structural properties of local government such as the percentage of college graduates or the proportion of the immigrants in one local community.

The second point regarding the combination of the different types of data is that we are having all information of Twitter users in South Korea. In fact, Twitter Inc. in South Korea provides us with such information and we have a deal with Cyram, which is specializing in social network analysis. They invented NetMiner, which is like UCINet or Pajek for the network analysis. Therefore what we do is to identify individual users for the social survey. We are trying to combine information from their text on the Twitter and the responses in actual survey, so that we can use the big data and survey data together.

Ronald Holmes
I would like to comment on the cross-national surveys.

We've quite a number of items that are run in cross-national surveys, and found out, for example, whether people are more populist than liberalist. We tried out this module developed by Americans and Europeans to test whether nationals of a particular country are more populist than liberalist. When we translated the questions, we had difficulty in translating the questions, because, the notion of diversity in the Philippines means "freedom". I don't know whether other Asians would understand that and whether we can translate it in such a way that it would more or less reflect our experiences.

In the Western world, when they talk about diversity they understand what it means. But in many Asian countries, the notion of diversity means different, which may be too abstract for us to get agreement of what it means. In cross-national surveys, we have to find a way to how to proliferate to it. Some of the questions in the survey may not necessarily correspond to each other. What we understand in the Philippines would not correspond to the one that Indonesian do, because we have different contexts and experiences.

Weidong Wang
I also want to give some example on how to combine different types of data.

The sample size of Chinese Education Panel Surveys is more than 20,000 and this is national representative sample. In this survey, we asked parents, teachers, and the principals of the students surveyed to cooperate so that we can combine the survey data with the student's hospital record data.

It was very difficult to get the consensus, but nearly 95% of students sent information about their hospitalization. We also combined student's questionnaire data with our psychometric test data including their personality, motivation and self-efficacy.

One of the purposes of combing the data in such a way was to prevent myopia, which is becoming more and more popular in China. In order to conduct a survey, I gave every student a wristband as gift. With this wristband, we can continually collect information about the movement and the location wherever they are. We also asked students to provide social media data in the WeChat or QQ to analyze their behavior.

Another example is the case of Chinese National Bank which has all the people's bank accounts. We did some research based on this data, randomly select samples and send the questionnaire to them. We deleted all the identification information but account information to maintain privacy.

Shigeto Sonoda
Thank you very much. Now let's start overall discussion and we would like to get questions from the floor. Are there any questions?

Chisako Masuo
My name is Chisako Masuo and I am from Kyushu University. I specialize in Chinese foreign policy, so I usually deal with qualitative data, not the quantitative one. Considering future possibility of collaboration with Chinese colleagues, I would like to ask Prof Wang two questions.

First question is, for the future international joint research, under what conditions do you think we can make Chinese government give permission for the foreign groups to collaborate with the Chinese team? Another question is how much collaboration or exchange of the information do you have with those people in party and government?

Weidong Wang
As to the first question, actually, it's very difficult. As Chinese know, when we want to conduct some international collaborative research, we must get some permission from governmental agencies. In our system some kind of license is effective when we apply for the permission, but we have never got any permission by now. There is a license hanging in the wall of my office at Renmin University of China, but we never used it successfully. Most importantly, we cannot get funding to those who are from other countries. So, if we want to get funding in China, try to fund those who share the common interest with you who might work with you. Then you can use such social network - that's my advice.

As to the second question, I think it's nearly impossible in China. Even for us it has become more and more difficult. So I don't know what we are having in the future, but I think the current government changed the situation too much.

Shigeto Sonoda
Any other questions?

Mie Ohba
My name is Mie Ohba from Tokyo University of Science. I've been studying the development of the regional institution in Asia, including ASEAN. So, I want to ask Professor Holmes. The first question is, do you have any international collaboration project to capture or construct the dataset with the research institutes or universities in the rest of the ASEAN member countries? Second question is, if you don't have any such international collaboration projects, do you know the situation in the Southeast Asian countries to construct the dataset of social sciences?

Ronald Holmes
As to the first question, we did a project based at Australian National University that looked at money politics, which included experimental surveys, too. The surveys were done in Malaysia by Merdeka Center, a survey organization; in Indonesia by LSI (Lembaga Survei Indonesia); in the Philippines by Pulse Asia. We were supposed to do it in Thailand, too, but a coup occurred. That's the type of collaboration.

We have not had collaboration among survey institutions in Southeast Asia largely because Southeast Asia is quite diverse. I will not mention the name of the ASEAN country, but I felt that while I was in that country, I was being followed - and I think I was -, but anyway I got home safely.

Social Weather Stations has been involved in international collaboration projects including Asian Barometer project in National Taiwan University, ISSP, World Values Survey, and electoral study survey as I mentioned earlier. I used to work with SWS, so they have continued collaboration with us. We have done some studies, but not necessarily with institutions in Southeast Asia, but more with multilateral institutions. ASEAN is the least cooperative organization in terms of data generation.

Shigeto Sonoda
One of the interesting findings of Asian Student Survey, whose second round survey was conducted in 2013 by undergraduate students of University of Tokyo and Waseda University, is that ASEAN students have more positive views toward US and Northeast Asia but they have relatively negative views toward neighboring countries.

Weidong Wang
I want to add some words about the international collaboration in China.

It must be easy for several large-scale social survey projects in China to provide international researchers with the opportunity for the open module. If you think some question is very important for the China studies, you just submit your proposal to the Chinese General Social Survey, China Family Panel Survey, or Chinese Education Panel Study. Just send proposal in English. If your proposal will be accepted, you can get the data. You do not need to pay for it. Chinese government will cover the fieldwork cost. That is the easiest way.

Shigeto Sonoda
Thank you for your wonderful information, Prof. Wang.

I believe we could have very good discussions. We are now taking a record and we are going to get the transcript so that those who are interested in the topic will be able to get some inspiration from today's discussion.

Now we will get concluded Kashiyama Seminar. Thank you very much for your attention and participation.