HITOTSUBASHI UNIVERSITY

Global COE Hi-Stat Discussion Paper Series 105

Research Unit for Statistical
and Empirical Analysis in Social Sciences (Hi-Stat)

# The Environment for Microdata Access in Japan: A Comparison with the United States and Britain and Future Issues

Akira Kawai
Shigeru Hirota
Tomohiko Inui

December 2009

Hi-Stat Discussion Paper

**Hi-Stat**
Institute of Economic Research
Hitotsubashi University
2-1 Naka, Kunitatchi Tokyo, 186-8601 Japan
http://gcoe.ier.hit-u.ac.jp

# The Environment for Microdata Access in Japan: A Comparison with the United States and Britain and Future Issues[1]

Akira KAWAI (Cabinet Office)
Shigeru HIROTA (Cabinet Office)
Tomohiko INUI (Cabinet Office)

Abstract

For most of the post-war period, Japan's administration of statistics was governed by the framework provided by the Statistics Act from 1947. However, because the Act remained largely unchanged since it was originally introduced, it increasingly failed to reflect important changes in economic and social circumstances over time, resulting in various problems, including with regard to the secondary use of various kinds of microdata. To help resolve these problems, the New Statistics Act was enacted in 2007 and came fully into force in April 2009. Among other things, the New Statistics Act provides for a substantial revision of the system of secondary data use.

An important element of this is a change in the basic philosophy underlying the legal framework from "statistics for the purpose of administration" to "statistics as an information resource for society." A central aim is ensuring the “usefulness” of public statistics, and regulations concerning the use of statistics, such as provisions for secondary use, were incorporated in the Act. One important change is that the system of approval by the Minister of Internal Affairs and Communications for secondary data use was abolished. Instead, secondary data use can now be directly approved by the survey implementer and procedures have been simplified, so in the new system

[1] This paper benefited greatly from valuable comments from the staff of the Statistics Commission of the Cabinet Office, particularly Director Hiroyuki Kitada, and from Sei Ueda, Deputy Director for Statistical Planning, Office of the Director-General for Policy Planning, Ministry of Internal Affairs and Communication. We would like to express our gratitude here. Needless to say, all remaining errors in this paper are the sole responsibility of the authors. The content and opinions are the personal views of the authors and do not in any way represent the official view of the Cabinet Office.

secondary data use now is considerably easier. Moreover, the New Statistics Act now allows for the provision of anonymized data and for custom tabulations for the purpose of academic research and higher education.

As a prerequisite for the promotion of secondary data use regulations with regard to users were also introduced. Other than the traditional duty of confidentiality, regulations concerning the appropriate management of questionnaire information and penal regulations for abuse of confidentiality were newly established.

In this way, the secondary use system in the New Statistics Act brings Japan in line with systems for accessing microdata found in Europe and the United States.

## 1 Introduction

In recent years, the use of microdata collected as part of official government surveys has come to play an increasingly important role in a wide range of academic fields. In the field of economics, for example, such data have made it possible to construct extensive panel data sets for a more rigorous analysis of a wide range of issues and from new perspectives. Concrete examples include empirical studies using firm-level or even establishment-level data examining the effects of foreign direct investment or international outsourcing on aspects such as domestic wage levels, skill upgrading, or spillover effects. Similarly, in the field of productivity analysis, there has been a dramatic increase in recent years in studies decomposing productivity using longitudinal micro-level datasets (see Bertelsman and Doms, 2007).

This means that both for academia and government, having access to microdata can be crucial for the conduct of leading-edge research both to advance our knowledge and for the purpose of policy formation. At the same time, however, providing access to microdata also brings with it a host of legal and practical problems, primarily related to issues concerning the confidentiality of information obtained from survey participants. Against this background, the purpose of this paper is to discuss the legal situation regarding the secondary use of questionnaire information from government surveys in Japan. Specifically, the aim here is to consider the role of the New Statistics Act introduced in 2007, issues related to the anonymization of data for secondary use, and the practicalities of providing onsite access to a wider range of researchers, and to compare these aspects with the situation in the United States and Britain.

The remainder of the paper is organized as follows. The next section provides an outline of the legal background regarding the production and use of government statistics in Japan. Special attention will be paid to secondary data use. Next, Section 3 outlines the situation concerning the production and provision of anonymized survey data for secondary use. In addition to that in Japan, the situation in the United States and the Britain will also be discussed. Section 4 then turns to onsite access to microdata, again comparing the situation in Japan with that in the United States and Britain. This is followed, in Section 5, by considering the framework for grasping the needs of statistics users and how to incorporate these into the production of statistics. Remaining tasks for further improving the situation regarding the use of statistics in

Japan are then discussed in Section 6, while Section 7 concludes.

## 2 Promotion of the Effective Use of Statistical Data

### 2.1 The New Statistics Act and Secondary Use

For most of the post-war period, Japan's administration of statistics was governed by the framework provided by the Statistics Act from 1947 (hereafter referred to as the Old Statistics Act) and the Statistical Reports Coordination Act from 1952. However, because these Acts remained largely unchanged since they were originally introduced, they increasingly failed to reflect important changes in economic and social circumstances over time, resulting in various problems, including with regard to the secondary use of statistics. Consequently, in 2007, the New Statistics Act was passed, which provides a comprehensive legal overhaul of, and integrates, the two former Acts. The New Statistics Act came into force in April 2009 and, among other things, provides for a substantial revision of the system of secondary data use.

To understand the revision, it is useful to briefly consider the regulations concerning secondary data use in the Old Statistics Act. In this context, it is important to note that the Old Statistics Act essentially was intended for regulating the production of survey statistics and largely failed to provide rules concerning their use. This is a reflection of the fact that at the time that the law was enacted, the only computers available were very large and expensive, methods of statistical analysis were much less advanced, and the main users of statistics were administrative organs. For this reason, the secondary use of survey data was considered as falling outside the original survey objective and therefore, in principle, prohibited, even for those implementing the surveys (Old Statistics Act, Article 15, Paragraph 1). In other words, the secondary use of data was treated as exceptional, and regulations concerning such use lacked particulars.

Specifically, with regard to questionnaires of designated statistical surveys, Article 15, Paragraph 2, of the Old Statistics Act states that secondary use was allowed if approved by the Minister of Internal Affairs and Communications. However, the Act did not stipulate any legal provisions regarding the requirements for such approval, the duties of users in handling questionnaires, or penal provisions for breaches of confidentiality. This clearly reflects the fact that at the time the Act was introduced, secondary use outside administrative organs was not envisioned.

Matters were simplified somewhat in the revision of the Old Statistics Act, when Article 15-2, Paragraph 2, was added. This allowed the secondary use of data from questionnaires of statistical surveys other than those for designated statistics (that is, notified statistics and approved statistics) based on the approval of those conducting the survey, provided that methodologies were employed that made it impossible to identify survey respondents. Yet, as before, there were no legal provisions regarding users.

Let us consider how the system of obtaining approval from the Minister of Internal Affairs and Communications operated in practice. The Old Statistics Act stipulated that approval for the use of secondary data was subject to the discretion of the Minister of Internal Affairs and Communications. In practice, however, approval procedures came in the form of a certain amount of paperwork that needed to be fulfilled. The basic criteria for approval were that the use of questionnaires did not break the confidentiality of survey participants and that it was of great public interest.
To give a few examples of purposes other than statistical purposes for which secondary data use was approved, these include the compilation of name registers, case studies, and dealing with violations of the Statistics Act. As for users, if public institutions such as administrative organs (but excluding public research institutions such as universities and research institutes) applied, they usually did not face any problems, because the users were public servants (including those deemed public servants). If research institutions such as universities and research institutes applied, it was necessary that either (1) they conducted joint research with an administrative organ; (2) they conducted research receiving assistance from an administrative organ; or (3) they provided a document from an administrative organ stating that the intended use is of great public interest. One reason why the range of users other than public servants was very narrow and essentially limited to those able to demonstrate that the intended use served the public interest is that in the case of public servants, the leaking of confidential information was subject to penal provisions under Article 19, Paragraph 2, of the Old Statistics Act as a breach of confidentiality, while for those other than civil servants, there were no penal regulations for the leaking of confidential information.

However, over the years, it became obvious that the Old Statistics Act was becoming increasingly outdated. Progress in information and communications technology as well as the development of new statistical methods have enabled researchers outside administrative organs to use statistical data in a large variety of ways, while at the same

time society’s statistical needs have become more diverse. Along with this, in countries other than Japan, there has been a growing emphasis not only on the production of statistics, but also on the use of statistics.

Against this background, as the use of microdata made significant advances internationally, the Old Statistics Act, which treats secondary data use as the exception rather than the rule, has increasingly come to be seen as an institutional constraint on academic research. That is to say, because scholars and researchers from outside the civil service were not granted access to questionnaire information to retabulate this and conduct analyses unless they fulfilled one of the three conditions mentioned earlier, there were growing concerns that it was impossible to conduct international-level research in Japan. Moreover, since even those implementing the surveys needed the approval of the Minister of Internal Affairs and Communications for secondary data use as it fell outside the original purpose of the collection of such statistics, statistical data were not used as effectively as they could be within administrative organs.

A key aim of the New Statistics Act therefore was to improve the situation regarding secondary data use. An important element of this is a change in the basic philosophy underlying the legal framework from "statistics for the purpose of administration" to "statistics as an information resource for society." Accordingly, a central motivation is ensuring the “usefulness” of public statistics, and as one aspect of this, regulations concerning the use of statistics, such as provisions for secondary data use, were incorporated in the Act. In fact, it was frankly recognized that Japan’s system for secondary data use was outdated and, based on the idea that what can be done in Europe and the United States can also be done in Japan, a new system was designed.

To start with, the system of approval by the Minister of Internal Affairs and Communications for secondary data use was abolished. Instead, secondary data use can now be directly approved by the survey implementer and procedures have been simplified, so in the new system secondary data use now is considerably easier. Concerning questionnaire information, as before, approval requires that the use is of great public interest. Use serving the public interest is defined as use by public institutions such as administrative organs, as well as cases determined in a Minister of Internal Affairs and Communications Directive as being of the same public interest as use by government institutions (New Statistics Act, Articles 32 and 33). The

above-mentioned cases (1) to (3) are set out in Article 9 of the Ordinance for the Enforcement of the Statistics Act.

Moreover, the New Statistics Act now allows for the provision of anonymized data and for custom tabulations[2] for the purpose of academic research and higher education (New Statistics Act, Articles 35 to 38). The Act recognizes that the results of such research are normally published in scholarly articles and hence returned to society, so that such use is approved on the basis that it "to a certain extent serves the public interest." It has therefore opened the way for researchers in universities and private research institutions to use government microdata. In the case of research by commercial organizations, too, use of the system is possible, as long as the main objective of the research is recognized to be academic and the results are published and thus returned to society, so that it fulfills the requirement of "to a certain extent serving the public interest."

As a prerequisite for the promotion of secondary data use, regulations with regard to users were also introduced. In addition to imposing confidentiality duties on users as in the past, the New Statistics Act newly stipulates users' duties with regard to the appropriate handling of questionnaire information and penal provisions for any breach of confidentiality.[3] As a result, the system for secondary data use put in place by the New Statistics Act is not very different from those in Europe or the United States. What is more, whereas in Britain and the United States the production and provision of anonymized data is based not on clearly stated legal rules but on legal interpretation, in Japan it now has a clear legal basis in the New Statistics Act.

### 2.2 The Role of the Statistics Commission

Generally speaking, secondary data use is now based on the assessment by the survey implementer. The exception is the production of anonymized data. The term "anonymized data" as used in the New Statistics Act means questionnaire information that is processed so that survey participants cannot be identified, for the purpose of

[2] Currently, under the custom tabulation scheme, those requesting a custom tabulation specify the content of table columns and rows and aggregate tables are then prepared for them. Another method currently being examined would be the use of computer programs such as statistical software packages by the survey implementer and the provision of the results to those requesting custom tabulations, but there are still a large number of technical challenges that need to be resolved.

[3] Penal provisions were put in place also for persons breaching confidentiality regulations outside of Japan (New Statistics Law, Article 62). However, secondary data use by foreign nationals poses various practical challenges and therefore will be dealt with in the future.

providing it for general use (New Statistics Act, Article 2, Paragraph 12). When an administrative organ intends to produce anonymized data pertaining to fundamental statistical surveys (which correspond to designated statistical surveys in the Old Statistics Act), it needs to hear the opinions of the Statistics Commission (New Statistics Act, Article 35, Paragraph 2).

## 3 The Environment for the Use of Anonymized Data

### 3.1 The current situation in Japan

In examining the current environment for the use of anonymized data in Japan, let us begin by considering the production of such data. At present, in 2009, the Statistics Bureau of the Ministry of Internal Affairs and Communications (MIC) produces and provides anonymized data only with regard to four fundamental statistical surveys (National Survey of Family Income and Expenditure, Survey on Time Use and Leisure Activities, Employment Status Survey, and Housing and Land Survey), while the situation with regard to other ministries and agencies is currently under review.[4] Meanwhile, based on Article 37 of the New Statistics Act, the Statistics Bureau has entrusted the provision of anonymized data and custom tabulation services to the National Statistics Center, an incorporated administrative agency.

The details of the production and provision of anonymized data are set out in the Basic Plan Concerning the Development of Official Statistics (hereafter, Basic Plan). The following is a summary of the steps outlined in the current Basic Plan for FY2009–FY2013, all of which are to be implemented in steps from 2009 onward:

- Each ministry and agency shall at the beginning of each fiscal year draw up and make public a fiscal year plan with regard to services and statistical surveys for secondary data use.
- Each ministry and agency shall implement paperwork based on a set of guidelines (the “Guidelines for Paperwork Concerning Anonymized Data and Custom Tabulations” set out in the “Ruling of the Director-General of Policy Planning of the Ministry of Internal Affairs and Communications”).

[4] For example, the Ministry of Health, Labour and Welfare and the Ministry of Economy, Trade and Industry are currently (FY2009) examining the relevant technical means for the production and provision of anonymized data.

- The MIC shall summarize the annual plans of ministries and agencies for secondary data use, the actual results of service provision, the number of cases where it was difficult to respond to filings, and the reasons for such difficulties, and publish an overview thereof and report to the Statistics Commission.
- Each ministry or agency shall seek to expand its services based on secondary data use needs and in line with the expansion of statistical resources.
- Each ministry or agency shall make the greatest efforts to secure the statistical resources for the purpose of adequately responding to needs for secondary data use.
- The MIC shall take necessary measures to maintain a system where the National Statistics Center becomes the repository entrusted with the provision of anonymized data and custom tabulations from each ministry or agency.

In this context, it is useful to consider the concept of anonymity in greater detail. In Japan, awareness of issues related to the protection of personal information grew considerably with the enactment of the Personal Information Protection Act and it seems highly unlikely that Japanese citizens would accept a concept as vague as that of “de facto anonymity” found in Germany, for example. For this reason, it was necessary to ensure that in providing secondary data use, no chance of identifying survey respondents remained. The challenge therefore was to, on the one hand, guarantee anonymization measures are in place to eliminate any chance of survey respondents being identified and, on the other, to avoid excessive measures in order to ensure that information remains useful for research.

The concept of identifiability of survey respondents in the New Statistics Act, which is defined in Article 2, Paragraph 12, is the same as the concept of personal identifiability in Article 5, item (i), of the Administrative Organs Information Disclosure Act and Article 2, Paragraph 2, of the Administrative Organs Personal Information Protection Act. In fact, the concept of personal identifiability in the Administrative Organs Personal Information Protection Act and the Statistics Act is taken from the Administrative Organs Information Disclosure Act, which was enacted earlier. Article 5, item (i), of the Administrative Organs Information Disclosure Act states that if an individual can be identified by cross-checking personal information in administrative documents with “other information,” this also constitutes individual identifiability. However, there are two conflicting interpretations of what is meant by “other information” in this context.

The first interpretation holds that this refers to information that any general member of society can obtain without exceptional effort. Consequently, the range of information constituting “other information” that could potentially allow identification through cross-checking is comparatively small, and those following this interpretation advocate a more laisser-faire stance with regard to making information available. In contrast, the second interpretation holds that “other information” also includes information that specific persons such as classmates, work colleagues, or neighbors can obtain. Those following this interpretation maintain that the range of “other information” that potentially allows personal identification is rather large, and advocates of this view are inclined to take a more restrictive view with regard to making information available. The draft of the Administrative Organs Information Disclosure Act took the standpoint of the second interpretation, while the Information Disclosure and Personal Information Protection Review Board adopted the former interpretation and administration in practice is conducted according to this former interpretation. The subsequent enactment of the Administrative Organs Personal Information Protection Act and the New Statistics Act were also based on the premise of the former view.

Consequently, the level of anonymization of data required by the New Statistics Act is that it makes it impossible for a general member of society to identify survey respondents. This, in fact, makes for a viable system of providing such data, since this level of anonymization means that information remains useful for research. On the other hand, given that survey respondents can potentially be identified by those with specific information, the use of such data is limited to academic research and higher education and comes with the obligation that users handle anonymized data with appropriate care. Although some think that it should be possible for anyone to freely download anonymized data, as is the case in some countries, this is problematic because the possibility of identification by persons who can obtain specific information remains. Another concern is that, even if anonymity were ensured, allowing unrestricted access to the recorded information of individual questionnaire responses could result in a loss of confidence by those participating in statistical surveys, so that providing such access does not appear desirable at the present point in time. In any case, since the use of anonymized data in academic research and higher education is permitted, as long as the research results are published and hence the fruits are returned to society, the lack of unrestricted access to such data is unlikely to be of great

inconvenience in practice.

Let us now turn to the measures taken in the anonymization of survey information. As mentioned earlier, the Statistics Bureau produces and provides anonymous data with regard to four fundamental statistical surveys (National Survey of Family Income and Expenditure, Survey on Time Use and Leisure Activities, Employment Status Survey, and Housing and Land Survey). Reflecting a prudent approach, these four fundamental statistical surveys were chosen to serve as first examples for the production and provision of anonymized data, avoiding statistical surveys where anonymization is comparatively difficult. That is to say, the view was that the construction of anonymized data is relatively difficult for surveys that survey the same participants again and again over a period of time, such as surveys focusing on establishments and firms or households and individuals, as well as census surveys.

The anonymization measures taken for the production of anonymized data from the four mentioned fundamental statistical surveys are listed below. However, it should be noted that because these four surveys serve as first examples, the measures taken were particularly strict. They are as follows:

- Not all original statistical survey records are used in the anonymized data, but instead a “thinning out” is applied to the original data (re-sampling of records).
- Moreover, records are arranged in a random sequence so that their sequence becomes meaningless (deletion of identifying information).
- Records containing numerical values that would allow identification are deleted (record deletion based on threshold values).
- As for extremely large values, an upper bound is created and such values are capped (top-coding).
- Classification categories are not very detailed; instead, broad categories are used (re-coding).

An important role in decisions on matters pertaining to anonymization is played by the Statistics Commission. Article 35, Paragraph 2, of the New Statistics Act states that with respect to the production of anonymized data from the questionnaire data of fundamental statistical surveys, the opinion of the Statistics Commission shall be heard. Consequently, the Statistics Commission deliberates on the sufficiency of anonymization measures from a specialist and technical viewpoint to ensure the

anonymization of data and the confidentiality of survey participants.

Moreover, the deliberations of the Statistics Commission also deal with technical issues to consider the usefulness of anonymized data and, in order to raise their usefulness as much as possible while ensuring anonymity, seeks to avoid excessive anonymization measures. In other words, the Statistics Commission deliberates on anonymization measures aiming to strike a balance between ensuring the anonymity and the usefulness of data.[5]

### 3.2 The current situation in the United States

Let us now examine the situation in the United States. We begin by looking at the provision of anonymized data by government institutions. Anonymized data of the American Community Survey in 2000 and the Census of Population and Housing for previous census years is provided by the Census Bureau of the Department of Commerce in the form of Public-Use Microdata Samples (PUMS). In U.S. census law, the principle that data shall not be published in a form that allows the identification of individuals or establishments is applied. Hence, anonymized data is provided per legal argumentum e contrario. There are no restrictions in terms of who can use the PUMS or the purpose they are used for. The pros and cons of the provision of individual pieces of microdata are decided by the Disclosure Review Board within the Bureau, which consists of employees of the Census Bureau. Under this scheme, two sets of data are available: the 1-Percent PUMS and the 5-Percent PUMS. The 5-Percent PUMS provide detailed regional characteristics, while the 1-Percent PUMS provide detailed item classifications. Using a specialized system called "DataFerret," PUMS can be accessed online.

[5] Concretely, the Statistics Commission determined that anonymization was excessive with regard to the following two points. The first concerns records of subsamples of the National Survey of Family Income and Expenditure. Here, it had been planned to delete from the anonymized data observations for households whose annual income, savings, and liabilities exceeded a certain value. However, it became necessary to retain such observations in the anonymized data and instead limit the information provided to the sum totals of annual income, etc., and apply anonymization measures such as top coding to households whose sum totals of annual income, etc., was high. Second, with regard to top coding of individual's age, the plan had been to set this to 75 years and above in the National Survey of Income and Expenditure and Housing and Land Survey, to 85 years and above in the Survey on Time Use and Leisure Activities, and to 80 years and above in the Employment Status Survey; however, due to the aging of the population and the importance of the analysis of the employment structure and family relationships using age, it became necessary to set this in all four surveys to 85 years and above.

The Statistics Commission found anonymization measures insufficient with regard to one point. Specifically, for the provision of information from the National Survey of Income and Expenditure, the Survey on Time Use and Leisure Activities, and the Employment Status Survey, it had been planned to geographically divide Japan into six areas, but in order to sufficiently ensure the anonymity of survey participants, it became necessary to distinguish only two groups, the "three metropolitan regions" and "other regions."

Apart from PUMS, the Census Bureau also operates a system for onsite access of detailed sample data, with use of such data being strictly limited to those with special sworn status (as well as being restricted to the place of access). (Business data are also available. They can be used onsite at the Center for Economic Studies and its network of Research Data Centers.) However, to obtain this status, users must satisfy certain conditions such as possessing knowledge that is useful for the activities of the Census Bureau and being employed by an institution providing services to the Census Bureau. The pros and cons of granting access are decided by the Disclosure Review Board.

In addition to the Census Bureau, microdata are also provided by the University of Michigan. Specifically, it provides microdata deposited with it by government bodies and private research institutions to institutions of higher education participating in the Inter-University Consortium for Political and Social Research (ICPSR). The anonymization of data is conducted at the source, i.e., by the institution depositing the data.

### 3.3 The current situation in the United Kingdom

Turning to the situation in Britain, we begin with the provision of anonymized data by government bodies. Anonymized data related to the results of the Census, Labour Force Survey, General Household Survey, etc., is provided by the Office for National Statistics. In UK census law, the principle that "individual census information" that makes it possible to identify individuals or households shall not be disclosed is applied. Hence, as in the United States, the provision of anonymized data is provided per legal argumentum e contrario. Depending on the data concerned (i.e., the particular statistical survey), restrictions are imposed concerning the purpose for which data can be used (i.e., academic or policy use) and who can use such data (i.e., researchers), and users need to sign a data access agreement. A research group set up before the production of census microdata examines and provides recommendations on the methodologies to be employed.

In addition to the ONS, two universities provide microdata. First, the UK's national data archive, the UK Data Archive (UKDA) is set up within the University of Essex and provides microdata deposited with it by government bodies and private research institutions through the Economic and Social Data Service (ESDS). To ensure the

protection of confidentiality, data collection deposit forms are examined by a committee.

Internet access to anonymized data is possible through the Internet authentication system "Athens," which requires users to obtain an ID and log in. In the case of use for non-commercial purposes, such as academic research, and for policy purposes, access is free of charge, but for commercial purposes, a charge is applied. In addition, for access to so-called "Special License" (SL) data (data that are not completely confidential but that make it possible to obtain detailed information) provided by the ONS, it is necessary to obtain a Special License from the ONS via the UKDA.

The second university in the UK providing microdata is the University of Manchester through its Cathie Marsh Centre for Census and Survey Research, which provides "Samples of Anonymised Records" (SARs) deposited by the ONS. If the data are used for academic research, users can download SARs free of charge after obtaining a user ID for the Athens authentication system and registering with the Census Registration Service. On the other hand, if the data are to be used for commercial purposes, users need to fill in and sign an end user license agreement, and a charge applies. Similarly, in the case of data use for policy purposes, there is also a charge. Two kinds of SARs are provided: a 2 percent (in some years 3 percent) Individual SAR and a 1 percent Household SAR. In addition, there is also the Small Area Microdata representing 5 percent of the population, which provides geographical detail at the local authority level but less individual detail than the Individual SAR.

The following five measures to protect confidentiality are in place and the disclosure risk through SARs is negligible:

- a low sampling ratio is used (individuals: 2 percent; households: 1 percent);
- data suppression is used (rare or characteristic observations are removed or merged);
- top coding or unification of categories (age, work hours, industry, occupation, education, etc.) is employed;
- regional information is limited (for individuals: limited to Government Office regions; for households: limited to England and Wales; no geographic breakdown for 2001 UK census)
- disclosure risk is assessed (using a model developed with funding received from the European Union).

In addition, Controlled Access Microdata Samples containing more detailed household and individual data can be used onsite within the facilities of the ONS with the permission of the ONS upon application.

## 4 The Environment for Onsite Access

In the preceding section, we talked about the production and provision of anonymized data. However, in panel surveys and statistical surveys of establishments and firms, for example, it is often extremely difficult to provide questionnaire information that is sufficiently anonymized yet remains useful in practice. In such cases, one possibility to provide access to this type of information is through so-called onsite use where researchers handle data within a facility that is strictly managed by the institution holding the survey information, and researchers only take out their research results.

However, efforts in Japan with regard to the provision of onsite access are still very limited, and what kind of practical set-up is appropriate is currently being examined by the Director-General for Policy Planning of the Ministry of Internal Affairs and Communications and the statistics departments of the different government agencies. In what follows, we discuss aspects related to onsite access in Japan and consider efforts in the United States and Britain.

### 4.1 The current situation in Japan

Onsite access in Japan is dealt with within the framework for the use and provision of questionnaire information set out in the New Statistics Act. However, due to budget and personnel constraints for the administration of such a system, there has been insufficient progress. A procedural framework that would grant access to a wide range of researchers has also not been established, and there is no provision of onsite access facilities to accommodate users living in different parts of the country. Against this background, a number of trials are currently taking place, and based on the outcome of these trials, the future course of action will be examined.

Let us begin our discussion of the details with the provisions of the Statistics Act. As mentioned above, the provision of questionnaire information outside the ministry or agency that conducted the survey is regulated in Article 33, the provision of custom

tabulations in Article 34, and the production of anonymized data in Articles 35 and 36. Because onsite access implies that outside researchers directly access questionnaire information that has not been anonymized, Article 33 applies. This stipulates that the provision of questionnaire information to anyone outside the ministry or agency that conducted the survey can only be for "the production of statistics or for statistical research" of similar public merit as that of public institutions. As stated earlier, cases of "similar public merit" require either that (1) they consist of joint research with an administrative organ; (2) involve research receiving assistance from an administrative organ; or (3) a document from an administrative organ, stating that the applicable use possesses public merit, is attached (Article 9, Ordinance for Enforcement of the Statistics Act). Thus, under the current legal system, onsite access requires joint research with an administrative organ or public assistance.

Current efforts for the provision of onsite access include the following. First, a scheme at the Statistical Research and Training Institute of the Ministry of Internal Affairs and Communications, under which researchers join research implemented at the Institute and are given the status of part-time national public servants. This then allows them to gain onsite access within the Institute to questionnaire information under the jurisdiction of the Statistics Bureau. This arrangement means that it is difficult for users to conduct research on a freely-chosen topic. On the other hand, the fact that the status of public servant is conferred on users and the duty of confidentiality imposed is an aspect in common with the practical set-up within the U.S. Census Bureau, as will be seen shortly, and this arrangement provides lessons for the way onsite access should be handled in the future.

The second scheme is the "Framework for the Use of Questionnaire Information Using an Onsite Access Facility" at the National Statistics Center, Incorporated Administrative Agency. Based on this framework, the National Statistics Center is entrusted with questionnaire information from administrative organs, provides affiliated academic research institutions with a facility for onsite access, and provides approved users with access to questionnaire data according to Article 33 of the Statistics Act. However, decisions regarding which questionnaire information is provided, and at which facility onsite access is allowed, are made at the discretion of each survey implementer.

At present, the only questionnaire information deposited at the National Statistics Center is that from statistics under the jurisdiction of the Statistics Bureau, and the

only affiliated academic institution is Hitotsubashi University. Thus, the expansion of both the data that can be used and facilities for onsite use remains an issue for the future.

### 4.2 The current situation in United States

In the United States, the major primary statistics are produced by the Bureau of Labor Statistics and the Census Bureau, and each of these provides facilities for onsite access.

(1) Bureau of Labor Statistics (BLS)[6]

Provides longitudinal employment data, price indexes, wage statistics, etc. (see Appendix 1). For access, the following two requirements must be satisfied:

(i) Researchers must either be a citizen of the United States, or be a citizen of an eligible country recognized by the State Department.

(ii) Researchers must either be enrolled as a student in an eligible educational institution or be employed by one of the following organizations: an institution of higher education; an eligible nonprofit organization; a state, local, or Indian tribal government; an associate of State or local public officials; or a Federally funded research and development center.

Data are available for research that is exclusively statistical. Once an application is approved, the researcher and his or her employer or educational institution must enter into a written legal agreement with the BLS. All work with confidential data must take place at the BLS national office in Washington, D.C. All printouts and storage media will be reviewed by BLS staff before they can be taken out to ensure that data confidentiality is protected. The use of the facility is free of charge.

(2) Census Bureau[7]

A large amount of questionnaire information from statistical surveys, such as each census, are available for access (see Appendix 2). For access, the following two requirements must be satisfied:

[6] The information here is largely taken from the BLS website (http://data.bls.gov/cgi-bin/print.pl/bls/blsresda.htm; accessed September 9, 2009) and from Kambayashi (2007).

[7] The information here is largely taken from the following website: http://www.ces.census.gov/index.php/ces/researchguidelines (accessed September 9, 2009).

(i) A research proposal must pass a review process (major criteria include scientific merit; benefit to programs the Census Bureau conducts under Title 13 U.S. Code; clear need for public data; feasibility; and risk of disclosure).
(ii) Researchers acquire Special Sworn Status (SSS), meaning they will be subject to punishment if they disclose confidential information. In addition, all SSS individuals must take annual training in the use of Title 13 data.

Access is provided at Research Data Centers (RDCs) at nine locations across the United States. The RDCs are jointly run by the Census Bureau and research institutions such as universities. Usage fees are in the region of US$15,000 per year, although the exact amount differs depending on the RDC. In some cases, use is free of charge if a researcher belongs to an RDC-related institution.

### 4.3 The current situation in the United Kingdom[8]

In the UK, onsite access to questionnaire information on establishments and firms is provided by the ONS through its Virtual Microdata Laboratory (VML). Data provided cover a wide spectrum, ranging from population data to employment and industry statistics (see Appendix 3). To gain access, researchers are required to submit a research application, which needs to be approved by the ONS, and contracts with the researcher are then drawn up and signed.

Onsite access is available at ONS offices in London, Newport, Titchfield, Glasgow, and Belfast. However, permanent support staff are available only in London and Newport and first-time users are encouraged to use these offices. All output to be taken out of the facility will be reviewed by support staff. The usage charge is £52/day, but this is normally waived for UK research institutions.

## 5 Understanding Statistical Needs

In order to ensure that the right statistics relevant for academic research and policy formulation are produced, it is important to understand the needs of statistics users from the various spheres. This section discusses current practices in Japan and also

[8] The information here is largely taken from the following website: http://www.ons.gov.uk/about/who-we-are/our-services/vml (accessed September 9, 2009).

briefly summarizes those in the United States and Britain.

## 5.1 The current situation in Japan

Because Japan employs a decentralized statistics system, the basic principle is that each ministry or agency with jurisdiction over statistics deals with seeking to understand the needs pertaining to the statistics it produces. Therefore, in the implementation of major statistical surveys, each ministry or agency solicits the views of users from administrative organs as well as academic experts with regard to survey items and tabulated items, regional breakdowns, publication frequency, etc., and also solicits opinions and requests from statistics users on the website where it provides the survey results. Moreover, when preparing statistical survey plans, anonymizing data, etc., ministries or agencies hold research meetings that include academic experts and take into account the needs of statistics users.

In the case of the MIC, which has jurisdiction over the overall coordination of government statistics, measures for understanding statistical needs are set out in its Basic Plan. The measures include, for example, activities on the internet to grasp the needs related to secondary data use, and the improvement of a wide range of statistics. By sharing the information and requests gathered in this way with ministries and agencies, the MIC supports the maintenance and provision of statistics by the various government departments.

The statistical needs of academics and other researchers are reflected in the administration of statistics through the deliberations of the Statistics Commission, which consists of 13 experts (New Statistics Act, Article 46, Paragraph 1, and Article 47, Paragraph 1). In the United States, the Committee on National Statistics (CNSTAT; discussed below) is an advisory body similarly made up of experts; however, in contrast with CNSTAT, the Statistics Commission legally is a third-party body and for this reason its clout vis-à-vis the government can be considered to be significant.

Under the Old Statistics Act, too, the needs of statistics users were reflected through deliberations on the Statistical Survey Plan by the Statistics Council, which was also made up of experts, and in this respect the situation has remained largely unchanged under the New Statistics Act. In addition, though, the fact that in the New Statistics Act deliberations on the draft for the Basic Plan fall under the jurisdiction of the Statistics

Commission has extremely large significance in reflecting the needs of statistics users. The reason is that consultations with the Minister of Internal Affairs and Communications on the draft for the Basic Plan do not take the form of consultations with regard to the pros and cons of specific proposals, but instead take the form of the concretization of contents being entrusted to the Statistics Commission, so that the Statistics Commission now has great latitude and can sufficiently incorporate the needs of statistics users. Furthermore, the Statistics Commission has been given the authority to follow-up issues with regard to the enforcement of the New Statistics Act, so that even after the Basic Plan has been set out, it can respond when new statistical needs arise as a result of changes in economic and social circumstances and ensure that the needs of statistics users are appropriately reflecting in the administration of statistics.

Another important avenue for grasping the needs of statistics users is the exchange of ideas between the Statistics Commission and academia. Thus, in addition government efforts to understand the statistical needs of users and to incorporate these in the administration of statistics, like in the UK and other countries, there are also examples of institutionalized meetings with users providing an opportunity for statistics producers and statistics users to exchange ideas, which are then considered in deliberations by the Statistics Council on the draft for the Basic Plan.

As a result of such an exchange of ideas, it was decided in the Basic Plan that the Statistics Commission should continuously monitor and act upon cross-ministerial statistics needs that individual ministries and agencies may be unaware of. Thus, the Statistics Commission will at all times consult with the relevant academic community and determine how to respond to statistical needs, focusing in particular on the establishment of a system that cuts across ministries and agencies for the improvement and creation of statistics. Moreover, it will put the results to practical use in the handling and provision of statistics of relevant ministries and agencies, in revisions of the Basic Plan, and in deliberations on items for consultation.

### 5.2 The current situation in the United States

The National Academy of Science (NAS) was established by the signature of President Lincoln in 1863 as a private nonprofit organization, and is charged with conducting surveys, investigating, experimenting and reporting upon any subject of science or art at the request of government. Under the NAS, the Committee on National Statistics

(CNSTAT) was established, the purpose of which is to "contribute to a better understanding of important national issues by working to improve the statistical methods and information on which public policy decisions are based."[9]

### 5.3 The current situation in the United Kingdom

The Government Statistical Service (GSS), an umbrella body that binds together statisticians from across government departments and organizations, seeks to grasp statistical needs through statistical advisory meetings and GSS advisory groups. Set up by the GSS, the statistical advisory meetings provide an opportunity for the public exchange of ideas between statistics producers and users. In addition, there are about 50 advisory groups that are privately organized.

## 6 Remaining Tasks for the Improvement of the Use of Statistics in Japan

The discussion so far has concentrated on the environment for the use of public statistics in Japan. Against this background, the purpose of this final section is to consider future issues and to give an example of recent research using secondary establishment- and firm-level data.

### 6.1 Future issues

Many of the issues to be addressed in the future have already been touched upon above. For example, it was mentioned that because of the difficulty of eliminating identifiability of participants of firm or establishment surveys, no anonymized data from such surveys are produced, not only in Japan, but also abroad. However, compared with large firms, which can be easily identified when they are classified by industry and region, it should be possible to anonymize data for small and medium enterprises, and the Ministry of Economy, Trade and Industry is consequently currently examining the production of such data.

Another issue for the future is to raise the usefulness of anonymized data. Partly because the production of anonymized data by the Statistics Bureau at present is a trial run, anonymization measures are rather strict and there is therefore still scope for

[9] CNSTAT homepage: http://www7.nationalacademies.org/cnstat/ (accessed September 9, 2009).

raising the usefulness of such data. At the same time, there is also room for improvement with regard to understanding the needs for anonymized data.

A further issue is the need to produce a greater variety of anonymized data from the same survey. For instance, the anonymized data of the four surveys of the Statistics Bureau at present are processed in 5-year brackets for those aged 15 years and over, but separately for each year of age for those under the age of 15. However, it has been pointed out that for the Employment Status Survey, anonymized data should also be produced by one-year age interval for those aged 15 and over, as well as broken down by various categories such as occupation, industry, etc. In order to put this into practice, research is necessary on the possibility of identification of survey participants through the matching of various kinds of anonymized data created from the same dataset.

Another challenge is to broaden the range of researchers that can access establishment and firm data. Because the anonymization of questionnaire information from establishment and firm data is difficult, there is little choice but to rely, as before, on the framework for the provision of questionnaire information set out in Article 33 of the Statistics Act. As mentioned earlier, in order for researchers to gain access to questionnaire information, this article stipulates that the use of such information by researchers must “to a certain extent serve the public interest.” Yet, if questionnaire information can be used only for joint research with a public institution or as part of research receiving assistance from an administrative organ such as a Grant in Aid for Scientific Research, then this will put those with little opportunity for participating in such research, such as young researchers and graduate students, at a great disadvantage, and this may stifle their very ambition for research. Consequently, taking the United States and Britain as an example, it would be desirable to introduce a review process for research projects and allow onsite data access to a wider range of researchers, provided that the research is recognized to be of academic merit and the appropriate duties and restrictions, such as confidentiality duties, are imposed. However, because a revision of the Statistics Act would be necessary for this, a thorough debate is required. However, if onsite access, which makes it possible to control the handling of questionnaire information quite strictly, can be made available to a wider range of researchers, this should become an important avenue for broadening access to microdata while maintaining confidentiality.

As for future issues with regard to onsite access in Japan, these include the expansion

of facilities for onsite access and the expansion of statistical surveys available. Concerning the first issue, facilities for onsite access, these are at present limited only to a couple of institutions such as the Statistical Research and Training Institute of the MIC, and in the long-term, taking also into account the convenience of researchers living in regional areas, it would be desirable to set up facilities for onsite access across the country.

Needless to say, it would be unrealistic, though, for each ministry or agency with jurisdiction over statistics to independently set up a nation-wide network of facilities for onsite access and the MIC would need to take the initiative to coordinate efforts to establish such a network. However, the Statistics Bureau of the MIC does not have external offices that could set up onsite facilities in regional areas. One possible solution therefore would be to link up with research institutions such as universities and set up Research Data Centers, following the example of the U.S. Census Bureau. The National Statistics Center already has a framework in place for creating facilities for onsite access at affiliated academic institutions, and it can be expected that the number of universities joining this framework will increase.

With regard to the second issue, that is, expanding the range of statistical surveys available for onsite access, it would probably be going too far to say that it would be desirable to make as many surveys available as possible. Because onsite access – even though it is controlled – permits direct access to data where there are concerns that individual economic entities such as firms and establishments are identified, it is essential to carefully consider the risk of identification and the implications that identification would have. In this respect, it is again possible to refer to overseas precedents, such as practices in the United States and Britain.

Another issue to be considered is collaboration with universities. It is, for example, imaginable that anonymized data is deposited with universities, and that such universities play the role of data archives, as in the United States and Britain. However, because this would require Statistics Act reforms, further investigation is necessary. To work toward a statistical data archive, the Basic Plan provides for the MIC to obtain the cooperation of ministries and agencies, the National Statistics Center, academia, etc., and to set up a committee charged with examining the particulars, such as how to maintain and operate a statistical data archive, the functions it should have, the range of data it should comprise, and how such data should be stored. The Basic Plan

requires the MIC to present the results of these efforts by FY2013.

Yet a further point to be considered is collaboration with academia more generally. Under the current set-up in Japan, the statistical needs of academics and other researchers find their way into the administration of statistics through the fact that experts are members of the Statistics Commission and of the study groups of the various ministries and agencies, and that they individually respond to hearings held by administrative organs. However, the only interaction this set-up provides is at the personal level between the experts and the administration, and there have been insufficient attempts at cooperating with the wider academic community with an interest in statistics. The Basic Plan recognizes this issue and aims for a greater exchange of ideas between the Statistics Commission and the academic community.

The challenge now is to concretize this exchange of ideas by deciding the practicalities of how and when, what topics to focus on, and who or what exactly the relevant academic community consists of. One possibility is to exchange ideas with regard to the reform of the Basic Plan, since in the process of deciding on the current Basic Plan, no particular exchange of ideas with the academic community took place. Thus, by exchanging ideas with the academic community both in preparations for deliberations by the Statistics Commission and in the deliberation process itself, it should be possible to more clearly take the needs of statistics users into account.

Finally, another issue is to better understand the needs of young researchers. It would be difficult to claim that, under the current set-up, the views of a wide range of users beyond veteran academics such as members of the Statistics Commission that regularly interact with administrative organs, are properly taken into account. Therefore, the Statistics Commission should also consider conducting hearings in order to grasp the needs of young researcher that use statistical data in practice.

### 6. 2 Examples of research involving secondary use of establishment and firm data

As stated above, it was only in April 2009 that the New Statistics Act came fully into force, and it can be conjectured that at present, there are few published studies using microdata applied for on the basis of the New Statistics Law. However, even under the Old Statistics Law, already a large number of studies using microdata have been

published, based on joint research with an administrative organ. A survey of recent studies using Japanese establishment- and firm-level data and focusing mainly on productivity research is provided by Matsuura, Hayakawa and Kato (2008). The survey suggests that the quality of empirical research has already improved substantially, and we expect that it will improve further as a result of greater access by more researchers to Japanese microdata under the new system provided for by the New Statistics Act

## 7. Conclusion

For most of the post-war period, Japan's administration of statistics was governed by the framework provided by the Statistics Act from 1947. However, this framework had become increasingly outdated as a result of social and economic changes and no longer was able to respond to data needs, including the secondary use of various kinds of microdata. To address this situation, the New Statistics Act was enacted in 2007 and came fully into force in 2009, providing, among other things, a substantial revision of the system of secondary data use.

An important element of this revision is a change in the basic philosophy underlying the legal framework from "statistics for the purpose of administration" to "statistics as an information resource for society." As shown here, ensuring the “usefulness” of public statistics represents a key focal point of this revision, and regulations concerning the use of statistics, such as provisions for secondary use, were incorporated in the Act. One important element was the abolition of the system of approval by the Minister of Internal Affairs and Communications for secondary data use. Instead, secondary data use can now be directly approved by the survey implementer and procedures have been simplified, so that in the new system secondary data use has become considerably easier. Moreover, the New Statistics Act now allows for the provision of anonymized data and for custom tabulations for the purpose of academic research and higher education.

It is still too early to judge the outcome of the change in the Statistics Act, and much remains to be done. However, as shown here, the New Statistics Act has removed many of the barriers to secondary data use in Japan and presents an important step in the direction of establishing a system of providing onsite access to anonymized data that is on par with what it found in countries such as the United States and Britain.

References:

Bartelsman, E.J. and M. Doms (2000), “Understanding Productivity: Lessons from Longitudinal Micro Datasets”, Journal of Economic Literature, Vol. 38, September.

Kambayashi, R. (2007) “Beikoku ni okeru Seifu Tokei Kohyo Chosa Kokai no Genkyo ni kan suru Chosa Hokoku: Beikoku Rodosha Tokeikyoku, Beikoku Census Kyoku oyobi Canada Tokeikyoku no Onsite Research wo Chushin ni [Onsite Research in the U.S. and Canada: Governmental Data Availability in North America],” *Hi-Stat Discussion Paper Series* No. 206, Hitotsubashi University.

Matsuura, T., K. Hayakawa and M. Kato (2008) “Microdata ni yoru Seisansei Bunseki no Kenkyu Doko – Sanyu, Taishutsu, Keizai no Globalization, Innovation, Seido Kaikaku no Eikyo wo Chushin ni [Research Trends in the Analysis of Productivity Using Microdata: Focusing on the Effects of Firm Entry and Exit, and Economic Globalization, Innovation and Institutional Reform],” *RIETI Policy Discussion Paper Series* 08-P-007, Research Institute of Economy, Trade and Industry.

Appendix 1 Questionnaire Information Available for Onsite Access at the U.S. Bureau of Labor Statistics

- *Data on employment and unemployment*
  - * National Longitudinal Survey of Youth 1979 (NLS79) Zip Code and Census Tract Files
  - * National Longitudinal Survey of Youth 1997 (NLS97) Zip Code and Census Tract File
  - * NLS79 School Surveys
  - * NLS97 School Surveys
  - * National Longitudinal Surveys Original Cohorts: Mature and Young Women, and Older and Young Men files with Geographic Variables
  - * 1993 and 1995 Surveys of Employer Provided Training
  - * Quarterly Census of Employment and Wages Program, and the associated Longitudinal Database of Establishments covered by State Unemployment Insurance Programs
  - * Occupational Employment Statistics Survey
  - * Job Openings and Labor Turnover Survey
  - * Mass Layoff Statistics
- *Data on prices and living conditions*
  - * Consumer Expenditure Surveys
  - * Producer Price Indexes
  - * Consumer Price Indexes
  - * International Price Program
- *Data on compensation and working conditions*
  - * National Compensation Survey and its predecessors, including:
    - Employment Cost Index
    - Employer Costs for Employee Compensation
    - Occupational Compensation Survey
    - Industry Wage Surveys
    - Area Wage Surveys
    - Employee Benefits Surveys

* Survey of Occupational Injuries and Illnesses

* Census of Fatal Occupational Injuries

## Appendix 2 Questionnaire Information Available for Onsite Access at the U.S. Census Bureau

AHRQ Medical Expenditure Panel Survey Extract
American Community Survey
American Housing Survey
Annual Capital Expenditures Survey
Annual Survey of Manufactures
Auxiliary Establishment Survey
Business Expenditures Survey
Business Register Bridge
Census of Construction Industries
Census of Finance, Insurance, and Real Estate
Census of Manufactures
Census of Mining
Census of Retail Trade
Census of Services
Census of Transportation, Communications, and Utilities
Census of Wholesale Trade
Commodity Flows Survey
Compustat-SSEL Bridge
Current Population Survey
Decennial Census Long Form Sample
Economic Census of Puerto Rico
Employer Characteristics File
Employment History File
Enterprise Summary Report
Exporter Database
Foreign Trade Data - Export
Foreign Trade Data - Import
Form 5500 Bridge File
Geocoded Address List
Individual Characteristics File
Integrated Longitudinal Business Database
Longitudinal Business Database
Manufacturing Energy Consumption Survey

Medical Expenditure Panel Survey - Insurance Component
National Center for Health Statistics Data Extract
National Employer Survey
National Longitudinal Survey
Ownership Change Database
Quarterly Financial Report
Quarterly Survey of Plant Capacity Utilization
Quarterly Workforce Indicators
Standard Statistical Establishment List - non Name and Address File
Standard Statistical Establishment Listing - Name & Address File
Survey of Business Owners
Survey of Income and Program Participation - Panels
Survey of Income and Program Participation - Longitudinal
Survey of Industrial Research and Development
Survey of Manufacturing Technology
Survey of Plant Capacity Utilization
Survey of Pollution Abatement Costs and Expenditures
Unit-to-Worker File

## Appendix 3 Questionnaire Information Available for Onsite Access Provided by the Office for National Statistics of the UK

Annual Inquiry into Foreign Direct Investment (AFDI)
Annual Population Survey (APS)
Annual Respondents Database (ARD)
Annual Survey of Hours and Earnings (ASHE)
Business Enterprise Research and Development (BERD)
Business Spending on Capital Items (BSCI)
Business Structure Database (BSD)
Capital Stock
Community Innovation Survey (CIS)
UK Innovation Survey
E-commerce Survey (ECOM)
International Trade in Services (ITIS)
Labour Force Survey (LFS)
Monthly Inquiry into Distributive and Service Sectors (MIDSS)
Monthly Production Inquiry (MPI)
New Earnings Survey (NES)
New Earnings Survey Panel (NESPD)
Products of the European Community (PRODCOM)
Quarterly Capital Expenditure Survey (QCES or CAPEX)
Workplace Employment Relations Survey (WERS)