Strategies of Promoting the Use of Survey Research Data Archive

by Meng-Li Yang[1]

SRDA

Abstract
The Survey Research Data Archive (SRDA) is the largest data archive in Taiwan and in Asia. It collects not only survey data in social sciences but also raw data of major government statistics. These archived data have made significant contributions to research. Data and remote access service are provided without charge. In addition, an English website along with the English version of the data and their metadata will be available by mid-2013. To improve the search efficiency and promote itself among domestic researchers, SRDA began to launch a series of projects around June of 2011. These include the revision of abstracts, the construction of new search functions, and the compiling and circulating of a power point concerning the use of SRDA. This paper documents the endeavors, reports the current progress, and reflects on the experiences learned from the developments.

Keywords: Asian survey data archive, archive management, archive development

Introduction

Data sharing is an important trend internationally. Researchers deposit their data to archives after the research project is completed, while others go to the archive to identify and use these secondary data to support different research interests. However, optimal data sharing requires continuous hard work and innovation. Data archives not only have to actively persuade data owners to deposit their used data, but also have to promote their services to potential users. Most importantly, while making efforts to secure data confidentiality, archives must make data as accessible and "transparent" as possible, so that researchers can find out if any data suit their needs as easily as possible. Such efforts increase the user-friendliness of the data and thus increase the possibility of data sharing.

In 2011, Survey Research Data Archive (SRDA, https://srda.sinica.edu.tw/) of the Center for Survey Research (http://survey.sinica.edu.tw/) of Academia Sinica (http://www.sinica.edu.tw/index.shtml) in Taiwan began a series of projects aimed at achieving such goals. In this paper I share our experience in conducting these projects. After an introduction to SRDA and its mission, I describe the challenges we faced and how those challenges were met. The paper concludes with a discussion of the experiences.

The Survey Research Data Archive

SRDA was established in 1994 by the Center for Survey Research (CSR) of Academia Sinica, and managed by the Data Division of CSR. It is the oldest and the largest survey data archive in Taiwan and Asia. It currently has almost 1400 members, who can be faculty, research staff and graduate students at universities, or researchers in government agencies and research institutes. While any visitor can review documentation and summary statistics online (available in 2012), members can download datasets directly from the website, request assistance from SRDA staff, and use remote access for secure data. These services are all provided without charge.

SRDA archives both the raw data collected by major government agencies for the production of important government statistics and data collected by academics. Each dataset released by SRDA is carefully cleaned and documented, and is released with detailed metadata. By the end of 2012, SRDA has released 423 datasets collected by the government agencies and 1,132 datasets collected by the academics, totaling 12 GB. These data are a very important resource for research and teaching. In 2012, members initiated a total of about 13800 downloads. Access to these data is responsible for a large number of publications in major national and international journals. Although we are still trying to build up the database of publications based on data archived in SRDA, according to records available now, data from only the six major survey projects in Taiwan archived in SRDA are the basis of 384 journal articles up to mid-2012.

SRDA continues to improve its services. In 2009, SRDA inaugurated on-line analysis service using the Networked Social Science Tools and Resources (Nesstar) software developed by the Norwegian Social Science Data Services (the service was seriously underutilized up to the end of 2011, though, because only several datasets were uploaded to Nesstar due to limitations to be specified later). To insure appropriate data security, the Information Security Management System (ISMS) protocols (based on ISO27001) were introduced in 2010. In April of 2011, SRDA, along with the Data Division of CSR that manages it, was certified by the British Standards Institution (BSI) (ISO 27001:2005). To enlarge the audience for its holdings, an English-language version of the SRDA website along with the datasets and their metadata will be online around mid-2013. Most major datasets should have English versions available by then.

The Data Division organizes activities and produces communications to promote SRDA holdings and services to scholars in Taiwan. The Division holds at least one workshop each year on important themes. These themes include skills for collecting and cleaning survey data, using important longitudinal survey data series, sampling methods, and advanced statistical analysis techniques. The Division also issues a bi-weekly newsletter and a monthly e-digest to announce SRDA's newly released data and activities.

Despite these successes, however, there was a sense that the data services needed to be more user-friendly and that information about SRDA should be disseminated more widely.

A Motivation for Need of Improvement

Some promotion strategies were already forming in June of 2011, but results from a survey to some extent confirmed the need to promote SRDA and its service.

In October 2011, the National Science Council (NSC) conducted a web survey[2] to solicit the opinions of scholars in selected fields. The survey target was scholars and researchers in humanities and social sciences who had submitted a grant proposal to the NSC in the previous five years (but who were not necessarily SRDA members).[3] To gauge researchers' use of SRDA, I took advantage of the opportunity to add several items[4] to the survey. The results of the SRDA-related items confirmed our original impressions. Of the 3019 respondents, 52.7% (1590 persons) had not heard of SRDA, and only 18.4% (556 persons) are or had been SRDA members. Among the 28.9% (873 persons) that had heard of SRDA but had never been a member, 616 had never even visited SRDA web site, and 257 did visit but did not apply for a membership. Among these 257 people , 17% said that information about datasets was insufficient for an effective search, and 10% similarly said that it was difficult to find needed datasets, although 64% had no data need. Among those who are or once were SRDA members but never used SRDA datasets for research (N=295), 17% said they did not know how to find what they needed and 16% said they could not find what they needed, whereas 52% did not had data need. The two most important messages from the survey were that 1) more than half of the researchers who might find SRDA valuable were not aware of its existence; and that 2) among those who tried to obtain data from SRDA, about 30% were frustrated with the process.

The messages reflected the difficult position SRDA was in. Although SRDA strives to promote its collection and services to researchers, it seemed that only groups that are already familiar with survey data or, more specifically, with SRDA, can benefit from the activities. For example, attendance at SRDA's survey data workshops was limited to those who could participate in person. From my own contacts with colleagues in the social sciences, many colleges in non-northern parts of Taiwan were not aware of SRDA. Although others wanted to receive additional training, CSR's limited resources forced us to refuse requests to hold on-site workshops at universities in other regions of Taiwan. Clearly, we needed to promote SRDA more actively.

Second, the survey results demonstrated that the site's search efficiency was poor. For many reasons, the search function within the original SRDA archive was rather old and inefficient. For example, data produced by government agencies require a separate user application process, whereas datasets produced by academics do not. The result is that each type of data requires a separate search. In addition, whereas the variable-level search of Nesstar is not functioning, the most powerful search in the original archive function searches only the abstract; the other search areas being the project name, the name of the PI, the serial number of the datasets, keywords, and the subject domains of the project. However, SRDA relied on data depositors to provide abstracts and keywords. Unfortunately, depositors are not always aware of how important the abstract is in archive search and retrieval functions. Poorly written or cursory abstracts minimized the effectiveness of SRDA's original search functionality. Therefore, except for searching within abstracts, the other search functions require the users to already know specifically what datasets they are looking for.

Even the on-line analysis function of Nesstar was seriously underutilized. By the end of 2011, only several academic datasets were uploaded to Nesstar because Nesstar does not keep records of people who make downloads but CSR needs such records. Government data were not considered for Nesstar at all, for the same reason that the use of government data requires further application.

Strategies

The survey results provided evidence that SRDA should improve search and discovery efficiency and also actively promote its services to researchers across the country, building on promotion efforts that began in June 2011. In sum, three strategies were aimed at improving search efficiency, and one was to promote SRDA among all potential users.

1 Improving the search efficiency

1.1 Revising abstracts

The first project launched was revising abstracts so that they included more information from questionnaires and accurately reflected the content of their datasets. The overall goal was to improve the effectiveness of searches. Each revised abstract should contain the purposes (and history if applicable) of the survey project, contents of the questionnaire, the survey mode, the survey period, the target population, the sampling frame, the sampling method, and the sample size. I asked all the Data Division members to review the project proposal/report and the questionnaire for such information, and to revise the abstracts from the view point of the dataset. This was done in early 2012 for 22 waves of a longitudinal survey projects, totaling 44 datasets.

Checking and editing the revisions proved to be more daunting than anticipated. In the beginning, I doubted the value of including only the title of the questionnaire sections. Ideally, concepts would be the most helpful for searching. However, the questionnaire of a social survey contains measures of all kinds of concepts. It is impossible to include them all in the abstracts. In addition, assigning concepts requires expertise in fields relevant to the goals of the survey, although staff members assisting with this project specialized in statistics. The result was to compromise and use only titles of the questionnaire sections

to describe the contents. Although this compromise may decrease the potential use of abstracts in improving search efficiency, highly detailed abstracts are just not feasible.

However, it is still important that abstracts contain all the other pieces of information, so that users quickly have a concise idea about a dataset by reading the abstract. Therefore, later in the middle of 2012, I recruited a doctoral student good at writing. I worked with him on revising abstracts for several longitudinal survey projects, after which he began to work independently.

1.2 Constructing a new search function

Members of the Division offered much better ideas. Around September of 2011, they suggested that we model our search function after the Survey Question Bank maintained by the UK Data Archive at the University of Essex.[5] The Question Bank has a variable-level search function. By using the DDI (Data Document Initiative) format, with which Nesstar is compliant, we can create a search function that allows users to find out the items of interest along with the data file by specifying key words in the items. This way, users can quickly find the exact data by specifying words/phrases of items that they need. As long as researchers know what items to find, they do not have to go through every possible dataset. A search function like Question Bank makes up for the deficiency of the original SRDA search features and accomplishes what searching in abstracts cannot achieve. So we decided to construct a Question Bank for SRDA.

For this new search function, I recommended that the government data should be also made within the search area. So they have to be put in the Nesstar. However, to decrease the risk of exposing any level of confidential information in on-line analysis, we allowed Nesstar to perform only univariate analysis for the government data. The programming work began around the end of 2011. A Division member undertook the system analysis (SA), and the programmer of the Division did all the programming tasks. During this time we also uploaded all datasets, government as well as academic, to Nesstar. In September of 2012, we put the Question Bank on line for service under the function name "Search by Item Contents" (http://140.109.171.171/bank/).

This new search function eliminates all the hassles of searching in the original SRDA and takes full advantage of Nesstar software. That is, by linking the original SRDA database with Nesstar, the search function searches in Nesstar the contents (variable labels and variable concepts) of government datasets and academic datasets at the same time and presents the results separately. Users are linked back to the original SRDA for downloading or requesting datasets. For Nesstar's on-line analysis functions, users can use all the functions for academic datasets, but only the univariate analysis for government datasets. More importantly, the new search function offers two types of search. The first type is called "search for datasets." Any dataset is listed that contains all the texts/concepts of variables entered by the user, whether or not these appear in the same variable. Users can also limit the search by specifying the range of years when the data were collected, the range of the sample size, keywords, words in abstracts, the project name and the name of the PI. The second type is "search for a specific variable," which is actually a method transplanted directly from the computer program used by the Data Division to construct the Concept Bank (explained later). Using this option, one can enter up to five phrases to identify a variable in mind. Any datasets that contain a variable which includes all the texts entered is listed. One powerful feature of the new search function is that results of each search method can be modified by either of the two methods. Finally, as we are also constructing an English version of the archived data, both English and Chinese versions of an identified dataset are always linked together. This way, researchers will be able to use the English translation of the data directly if they wish to submit the analysis to an international journal. The English version of the SRDA website will also have this search function with all the features available, where English texts and English concepts are used for searching. The features of the new search function are summarized in Table 1.

	Search for a specific item	**Search for a dataset**
Type of search term	• variable text	• variable text • variable concept
Maximum number of terms	5	3
Applicable search restrictions	None	• Collection year range • Lower limit of sample size • Name of PI • Name of project • Words in abstract • Project keywords
Logic of search	Intersection：only variables containing all search terms are displayed	Union：datasets meeting the search criteria and containing all the search terms are displayed; search terms do not necessarily appear in an item.
Search further?	Yes, and can use the "Search for a dataset" to do further search.	Yes, and can use the "Search for a specific item" to do further search.

Table 1. The two types of search in the "Search by Item Contents" function

1.3 Constructing a new item-level search option—the Concept Bank

The Division started to develop a "Concept Bank" in 2010 but abandoned the project in early 2011, before I became the advising researcher of the Data Division. The idea of a "Concept Bank" is to assign concept(s) to every variable for the archived data, so that users can also use concepts to search for variables. However, in 2010 the Division did this by translating an English thesaurus for the social sciences to Chinese, a "top-down strategy." When this was almost done, they met with three obstacles. First, they found concepts that are not applicable to Taiwan's situation and vice versa. Second, there are concepts that seem to have more than one translation. Third, they could not find resources (expertise) to assign these concepts to items.

While I believed in the value of building the Concept Bank, I thought the top-down strategy was not efficient. Instead, I proposed a bottom-up strategy by asking experts to assign concepts for variables, in both English and Chinese. My idea was that the structure of concepts that a thesaurus offers may not be essential when used in variable-level searching. Especially, compared to the bottom-up strategy, the top-down strategy may require a great deal more manual labor—to link the concept for each variable back to the thesaurus— in addition to the expertise required for the assignment. Even asking experts to assign concepts from the thesaurus had limitations, and, after all, the thesaurus did not always incorporate concepts unique to Taiwan's situation. Also, concepts that do not apply to Taiwan's situation should be no concern at all because the archived data would not contain such concepts. The problem of one single concept corresponding to more than one translation does not need to be a concern either, since experts may know most of the translations, and future users should also know the several Chinese translations of a certain English concept and vice versa.

The Division's greatest concern for the bottom-up strategy, and also the reason why it adopted the top-down strategy before, was that different scholars are likely to define different concepts for identical variables, which would diminish the bank's value for searching. Anticipating this problem, I proposed designing a computer program to check for the inconsistency. People working on the project can use the program to check if variables that are supposed to have identical meanings do have identical concepts, and if not, they can output the concepts along with the variables and have the concepts revised by some other experts. Furthermore, I suggested that once we have concepts for a variable, we can save time and resources by assigning the same concepts to variables almost identical except for small differences in non-substantive words.

In short, we needed a computer program that allowed us to input and output concepts to Nesstar, and also to identify variables with identical substantive meanings except for some non-substantive words. Such a function was quickly designed and programmed by Division members. This program allows one to specify up to five phrases to locate a variable. The staff member uses it to identify variables that contain these phrases, check the concepts for consistency, and, if necessary, output them for revision, and, afterwards, input the revised concepts. When consistency is assured, the staff member uses the program to assign the processed concepts to all the other variables that are identical in meaning. I am responsible for revising concepts for items with inconsistent concepts. My principle of doing the revision is to include all the concepts unless they are obviously wrong. After all, concepts can be very specific or very general; including them all may serve researchers' different needs.

Thanks to resources from NSC, we invited 45 scholars to assign concepts for 45 surveys in 2012. In the first round of the invitation, we selected studies of a wide range of topics from several longitudinal projects. Each topic was represented only by one study. For example, only the most current one, rather than all, of the surveys on political science was sent for concept assignment. The same was done on topics of family studies, citizenship, secondary school students, teachers, etc.

Nevertheless, even surveys supposedly on different topics have many overlapping variables, and there is rather low consistency among the concepts assigned to identical variables, as the Division has expected. Further, the timeliness with which scholars returned the completed material varied widely. Consequently, even though we completed consistency check for variables within several studies, those received later introduced more inconsistencies. To accommodate this problem, we decided not to check until all the studies of the same project are returned. And then, after we have checked the consistency for variables in all the studies that were sent out, we can start to assign concepts to identical items in all the other datasets. By the end of 2012, we had completed the assignment for the 45 datasets of one longitudinal project (the goal in the grant proposal), and several other surveys of different series on a variety of themes. The concepts are put in Nesstar for use in the new search function. We are still waiting for more scholars to send back their work so that we can resume the consistency check. In 2013, we obtain additional funding from the NSC to invite scholars to work on additional surveys with different themes. Learning from the experience, we will be working with a smaller number of studies as we expect the load of consistency checking will increase as more studies are assigned concepts.

2Actively promoting SRDA across the country

As mentioned earlier, the 2011 survey results indicated that many potential users still did not know about SRDA and efforts to promote the archive to researchers were needed. Early in 2012, I decided to put together a Power Point program that introduces SRDA and also contain some examples of how archived survey data could be applied to research projects. The introduction of SRDA itself was easily completed, but the demonstrations had to be designed from scratch.

I chose to demonstrate the usefulness of survey data in two ways. The most obvious one is to point out research articles that analyze survey data. I wanted to focus on articles that are easily found in the internet and would encourage use of SRDA datasets. Since the Data Division did not have a spare hand for such a job, I asked my own part-time assistant to find such articles in TSSCI (Taiwan Social Science Citation Index) journals, and to compile an abstract for each. I spent a significant amount of time revising the abstracts. However, it became clear when the abstracts were inserted in the Power Point file that much of this effort was unnecessary. Because the Power Point file is for self-viewing, an article is easier to understand if it is discussed on one slide. However, a long abstract is too long to fit into one slide. The result was that I shortened the abstracts into research questions for each of the studies (22 studies) included in the Power Point file.

Another way of highlighting the potential value of less well-known datasets was to write up some analysis using the data to answer simple research questions. This strategy is adapted from the ICPSR On Line Learning Center (http://www.icpsr.umich.edu/icpsrweb/OLC/). Another part-time assistant of mine (a doctoral student) wrote the analyses. From those, I selected seven works for the Power Point file. I ran into the same problem as in the abstract case, and had to shorten the

complete analysis reports to include only the research questions, the title of the data, and a short description of the results.

The Power Point file was completed in January of 2013. Although we would have preferred to complete it earlier, the delay allowed us to include a brief introduction of the new search function (Question Bank and Concept Bank). The file was sent via email to all college professors in humanities and social sciences across the country in March 2013. In addition to informing the professors of SRDA, we suggested in the cover letter that they show the file to students in class. We hope that this will encourage more students and researchers to use the archived data

Conclusion

Since June of 2011, SRDA has been engaged in such developments to promote the likelihood of data sharing. It has completed a variable-level search function with variable concepts as a new search option, and compiled a power point file to promote its use. The other two projects, those of abstract revision and concept assigning, are still going on. To write a good abstract turned out to be more difficult than originally thought. As we can give only section titles, rather than major concepts, as the contents of a study, the potential value of abstracts for search may be reduced, unless there is a clear description of the theories or purposes to be tested by the study. Nonetheless, as we have a variable-level search function, users probably do not need to rely much on abstracts to find data. The assignment of concepts is the most resource consuming because it requires scholars' contributions as well as staff members' continuous checking for consistency. However, the construction of concepts has to continue if it is to contribute to the search efficiency. Concepts are valuable not only in searching for data but also in designing questionnaires when it is necessary to include a measurable theoretical concept.

During all this time, as an advising researcher to the Data Division, I have voluntarily involved myself in the developments, sometimes even using my own resource. Whereas the Division focuses on their routine tasks of data cleaning most of the time, I work closely with two or three of the members, who are more skillful in designing and programming. I regularly enquire about the details and progress of the projects and hold discussions, to make sure the projects are in the right track or to seek solutions to problems. When projects are in a good preliminary shape, ideas are also solicited from the Division or the CSR, which results in more improvements. Such close supervision had helped with the construction of Question Bank in two different stages.

From the experience, I realize the importance of organization and of the leader's active involvement when the business is just developing. To ask a researcher to oversee an archive will probably lead the archive nowhere because the researcher cannot pay too much attention. Therefore, for an archive to develop, it is important to have someone with research experience as its own director. A full-time director will be able to devote all the attention to the archive. The director will be able to not only learn about researchers' needs, learn about development policies and strategies from other archives, but also carefully plan for projects, and supervise closely the progress of the projects.

It is also important to equip the director with a team with various skills. For example, skills such as project designing, computer programming, formal document writing, in addition to data processing skills and data preservation knowledge and techniques, are necessary in the above projects. Without these skills, it is very difficult, if not impossible, for an archive to implement improvement projects. However, such people need substantial orientation and an environment that encourages skill development and innovation. After all, working for data is a very special application of their non-statistical skills. Without a nurturing environment, such people may soon feel frustrated and leave the organization. I myself actually had spent some time and the other two members also spent time listening and talking with such people, so that they felt they had someone (if not all) in the Division to rely on when frustrated. From the experience, I found that designating a mentor for them not only helped keep them in the organization but also enhanced their performance. The mentor does not have to be as skillful as the new colleague in the specialized area. The mentor just needs to be kind enough to be willing to help a completely new learner and to provide information on matters that are related to where the skill is to be applied.

NOTES

1. Center for Survey Research, Research Center of Humanities and Social Sciences, Academia Sinica. Address: 128 Sec.2 Academia Road, Nankang Taipei Taiwan 11529. Contact via email: mengliya@gate.sinica.edu.tw.

This is an expanded version of a paper presented on June 7, 2012 at the IASSIST 38th Annual Conference in Washington, DC.,

2. The web survey was implemented by CSR by sending an invitation via email to the researchers, email addresses being provided by the NSC. The email gave URL links and asking the receiver to answer survey questions on the web. There were three follow-up emails for people who did not respond to the survey. .
3. Since the NSC is the most important, if not the only, agency that supports academic research, these researchers may be considered constituting almost all of the scholars in Taiwan that do research.
4. The items used to gauge about use of SRDA are as follows. The first number in the parentheses following each response option is the frequency that chose the option. The percentage is the percentage that these people account for of the total number of respondents to the question

1. Are you currently an SRDA member? (N=3019)
(1) Yes, I am. (Go To Q2) (365, 12.1%)
(2) No. I was before, but the membership is not valid now. (Go To Q2) (191, 6.3%)
(3) No, but I heard of SRDA and that it provides free access to data. (Go To Q3) (873, 28.9%)
(4) No, I have never heard of SRDA. (1590, 52.7%)

(For those who are or were a member) (N= 556 =365+191)
2. Have you ever used data archived in SRDA for research?
(1) Yes. (261, 46.9%)
(2) No. (Go To Q2-1) (295, 53.1%)

2-1. What is the reason that you did not use data from SRDA for research?
(N=295, those who answered "No" to Q2)
(1) I do not have data need. (153, 51.9%)
(2) I don't know how to find the data I need. (50, 16.9%)
(3) I cannot find the data for my research. (46, 15.6%)
(4) I downloaded some data before but then I found that they did not fit my research purpose. (38, 12.9%)
(5) Others. (8, 2.7%)

(For those who heard of SRDA before, N=873)
3. Have you ever visited SRDA website?
(1) Yes, I did. (Go To Q3-1) (257, 29.4%)
(2) No, I did not. (Stop) (616, 70.6%)

3-1. Why didn't you apply for an SRDA membership? (multiple choice)
(1) I do not have data need. (164, 63.8%)
(2) The application procedures for a membership require too much. (38, 14.8%)
(3) The amount of data archived is not large enough. (23, 8.9%)
(4) The information provided on line is not sufficient enough to find data easily. (43, 16.7%)
(5) The interface on line is not easy to use to find data. (26, 10.1%)
(6) The application procedures for using the data I need (government data or secure data) require too much. (58, 22.6%)
(7) I cannot find data that meet my needs. (45, 17.5%)
(8) Others. (8, 3.1%)

5.http://surveynet.ac.uk/sqb/