International Documents Roundup

International Survey Data: Challenges and Strategies for Collection Development.

DttP: A Quarterly Journal Of Government Information Practice and Perspective, 36 (1), 12-16, 2008.

James Church

As a “hybrid” international documents librarian I liaise directly with our economics department working closely with faculty and students. One of the items graduate students request most frequently from me is survey data from developing countries. In the not-so-distant past, students had to move mountains to acquire data sets like these. Ten years ago, it would have been almost out of the question for a graduate student to approach a librarian and request household survey data from Brazil or enterprise data from India. But thanks to the Internet, students can easily find descriptions of international surveys online, and sometimes the data files themselves. If the data costs money, students are coming to librarians and asking for them.

Since survey data has not been traditionally collected by most libraries it is not something most government information librarians feel comfortable dealing with, and for good reason. It often costs hundreds or thousands of dollars; can be fiendishly difficult to acquire; and since the data is often confidential, containing the names of individuals or firms, often necessitates difficult contract negotiations and other legal hassles. In addition, survey data, or "microdata," differs markedly from national level or "aggregate" data in that it is typically composed of raw data that needs to be interpreted using documentation (e.g., a code book) or analyzed using statistical application software, such as STATA or SPSS. But should we neglect this issue in an era of rising user expectations and a burgeoning interest in the global economy? In my view, in an age

where this information is easily accessible in digital format, we ignore acquisition of this form of government information at our peril. Librarians need to develop innovative collection development models that meet the research needs of our communities. Such models will require mediation and funding from libraries.

Aggregate Data vs. Microdata

Aggregate data, for the sake of this article, can be defined as a data total created from smaller units. For instance, the population of a country is an aggregate of the populations of the cities and rural areas from that county. Statistical agencies like the World Bank and the U.S. Census Bureau often collect smaller data units from surveys of households or firms and then aggregate them for publication in the form of statistical tables. Aggregate level data is sometimes referred to simply as "statistics"— the means, ranges, and other aggregate descriptors of the underlying microdata. Examples of statistics in both the international and domestic arenas come readily to mind, including most of the statistical tables found in the *Statistical Abstract of the United States* and much of the national data found in international databases like World Development Indicators and Source OECD.

In contrast, microdata files contain information on individuals, firms, or other smaller or discrete units of a population. Microdata is compiled from surveys created by researchers at think tanks, universities, and *governments*. The United States Census Bureau conducts several surveys with which most government information librarians are familiar—for example, the Current Population Survey, County Business Patterns, and the American Community Survey. To protect the confidentiality of the participants, individual names of households or firms are removed or

"anonymized" and the data documentation is relatively easy to use. A quick search in the Intra-University Consortium of Social & Political Research (ICPSR) database will uncover many such surveys, and many others are freely available via the U.S. Census Web site and the sites of other US government agencies.[1] Recently, ICPSR created a new division called the International Data Resource Center (IDRC) that acts as a clearinghouse for its international data.[2] If you are like most other government librarians I know, you will quickly run to your data librarian for help downloading and manipulating these files.

The New Online Documentation

The problem for international documents librarians and librarians who deal with development economics is that many countries do not belong to ICPSR or any other consortium -- or if they do, our universities are not members. These data files are quite literally "all over the map," and until recently, they were almost impossible to acquire. But in recent years, new tools (not to mention Google) have brought these data sets to the forefront. One of the best examples I know of is a site developed by the International Household Survey Network (ISHN), an organization "seeking to improve the availability, quality and use of survey data in developing countries" (www.internationalsurveynetwork.org/home/). Current members of the network include eighteen International Governmental Organizations, including the World Bank, the United Nations Statistics Division, UNICEF, and the International Labour Organization.[3] The chief area of interest for librarians is, naturally, the catalog, which allows for browsing for surveys by country and subject. For example, a country search for India retrieves 71 national surveys, which can be broken down into labour force surveys, living standards measurement surveys, demographic and health surveys, and more. The metadata provided is helpful, and includes the names and

addresses of the producers, coverage, method, sample size, an abstract, and dates. If reports on the results and questionnaires are available, links are usually provided, as are links to the sites of the survey producers. The ISHN is not the only site sites providing information of this kind. Other international survey databanks include the Demographic Health Surveys (which features an online tool that allows users to select countries and indicators to create customized tables), the World Bank Living Standards Measurement Surveys (LSMS), and the UNICEF Multiple Indicator Cluster Surveys.

Aside from the usual problems with these gateway sites (broken links, invalid e-mail addresses, etc.), one issue should be obvious. . All these reports, questionnaires, and other metadata are useful, but most of the time, students don't just want the "statistics" or the results -- they want the microdata for their own purposes which allow researchers to perform cross-tabulations, and which can lead to the discovery of facts and observations other than those for which the survey was originally intended. This is where the trouble starts. Some government agencies make the data freely available online in anonymized form. ISHN is urging governments to do this and offers specific principles and guidelines for anonymization.[4] But in many other instances, the information is not free, or it has not been anonymized. The user (typically a broke and hapless graduate student) is now confronted with a host of obstacles.

Financial Considerations

Unfortunately, many survey producers have discovered that people are willing to pay for this information. Surveys can cost thousands of dollars to prepare and conduct, so this is reasonable. But the really large surveys, particularly longitudinal surveys (surveys conducted over time) and

panel surveys (surveys studying the same group of people), can be extremely expensive. Some of the surveys for which I have received requests cost hundreds of dollars, which, for a research library, is not a huge obstacle. But others costs thousands of dollars: the most expensive I have been asked about to date is the European Union Labour Force Survey (EU LFS), which costs 8,000 euros to obtain. I have also received requests for floating population surveys from Shanghai, household surveys from Brazil, enterprise surveys from India, and longitudinal monitoring surveys from Russia. My home institution, UC Berkeley, has a higher concentration of students doing development economics research than most, but we are certainly not alone. Princeton University has done an exemplary job of acquiring and providing international survey data to students and faculty. The Princeton Data and Statistical Services (DSS) site (dss.princeton.edu/cgi-bin/dataresources/guides.cgi) is a model for subject and country access to this information, offering guidance on international surveys across a wide range of subjects and countries. Some of these data are free or simply require registration, some are available via ICPSR/IDRC, and some are exclusively for Princeton. I know my colleagues at Stanford University are getting these requests and struggling with the same issues. Quite simply, international survey data is in demand. The issues of globalization and international political economy are pressing, and research conducted using microdata can offer insights not obtainable from the aggregate data.

Privacy, Contracts, and Sharing

While the costs of obtaining this data are high, the problem is not insurmountable if libraries and consortiums such as ICPSR and ISHN begin to engage in strategic collection development policies and best practices. But this is not the most pressing issue; the main problem is legal,

especially for state universities hampered by a quagmire of government regulations. As noted previously, some surveys contain the names of households, individuals, or firms. In some cases, the data has been anonymized, but in the developing world, this presents an additional cost, and is sometimes not done. Survey producers wish to protect the identities of their respondents, and often require researchers to sign confidentiality agreements before agreeing to provide data to a researcher.[5]

UNDERTAKING[5]

I, Dr./Mr./Ms…….....son/daughter/wife of ….....
Resident of ……. (full address)
and presently working as ……..in the ……..having obtaining the data as detailed below:

Round No. ……
Schedule No……
Subject of enquiry……..
For the purpose of ……..
hereby undertake to comply with the following terms and conditions:

(i) The confidentiality of the unit level data will be maintained and adequate precautions would be taken for not disclosing the identity of the units directly or indirectly.

(ii) The data will be used only for statistical research and analysis and not for any other purpose.

(iii) The data obtained will not be passed on either wholly or partially with or without profit to any other data user or disseminator of data with or without commercial purpose.

(iv) The data user shall acknowledge the data source in the research output. The research outputs along with the short summary of conclusions would be made available (to the agency) in the form of hard copy or on electronic media free of cost whenever requested.

Signature……..
Date…… Name ……..

There are a number of thorny problems with this. For a private researcher to sign this kind of agreement is not problematic. These are the users for whom this data is primarily intended. But

most libraries have never been involved in this kind of legal contract, and are likely to be incredulous. The first issue is that the library would need to store such personal data in a secure location in order to prevent "Joe Public" from walking in off the street and discovering the sexual histories of individuals in Russia or the names and wages of textile workers in Mumbai. The second is that, in order to ensure confidentiality, the library would be forced to play the intermediary between an individual and a foreign government. If there is a breach of confidentiality, who would be liable? (My limited sense of contract law tells me that the library could sign a disclaimer and place all responsibility on the user, but apparently it is not that simple). The third issue, and potentially the most problematic, is that some statistical agencies want to see the results of research conducted using their data. Naturally. The data was compiled to solve pressing economic and policy issues in their country, and if someone uncovers a potential solution, they want to know about it. Unfortunately, experts who work with survey data at the World Bank inform me that student compliance with this stipulation is approximately zero. Unless these issues are resolved, most libraries will not get into this business, and graduate students will remain frustrated by glimpses of microdata (microdata metadata) that they can read about, but not use.

A Potential Solution

I have been having conversations with colleagues about this problem. The easiest and perhaps the wisest approach would be to do nothing, and advise graduate students to go abroad to get the data (not a bad idea, since government bureaucrats have a habit of ignoring e-mails, phone calls, and faxes). This could certainly be a wise strategy when the survey documentation is indecipherable without assistance from a cryptographer. A second approach is to suggest that

students secure funding to obtain the data themselves via a grant or departmental assistance. This certainly is done, and often with excellent results. However, both of these approaches cut the library out of the picture as an information provider and place the burden on the user. That is probably not a good idea in the current academic climate. In my view, the library profession is under siege from a number of powerful economic and societal forces. Most government information will be available exclusively in digital format soon, if it is not already. In a world of e-libraries and Googlezons, our role as catalogers and even selectors of government information is diminishing, and in order to remain relevant we need to create new niches for ourselves. Providing the expertise and funding needed to acquire international surveys is potentially one such niche.

It may be time to consider a more user-driven collection development policy. Libraries have been buying books and serials for years, and until recently have not been paying extraordinarily close attention to usage before making purchases. If you select as I do, you probably purchase a monograph based on the academic appeal of the topic or author and the quality of the publisher (not to mention the price). But the fact remains that we buy many books, documents, and microfiche that few people use. In today's fiscal environment, it seems misguided to spend thousands of dollars buying items students have not specifically requested while neglecting to purchase data that could lead to the publication of a doctoral thesis. This may be why so many economics students and faculty whom I have spoken to do not see libraries as meeting their needs. I'll never forget the time I ran into a famous economist at a local café. His remark ("The library? Oh yeah, I've been there once,") made me cringe.

Change is possible, but we need to rethink our mission as librarians. It clearly does not make sense for us to begin haphazardly acquiring expensive international surveys with poor data documentation and absurd legal provisions. Instead, I suggest careful consultation with faculty and graduate students and a thorough negotiation process with survey producers. Then we can begin to collaborate with consortiums such as ICPSR and make strategic purchases of data sets. We may need to hire legal specialists to negotiate with foreign governments to handle the questions of confidentiality, venue, and publication rights. As time goes on, this will become easier. But the time to begin is now.

References

1. The ICPSR (www.icpsr.umich.edu/) is the world's largest archive of digital social science data and a good source for international survey data. Membership, however, is concentrated in western countries. There are several other national and international numeric data archives (see www.dialogical.net/socialsciences/directories.html for a decent list) but none of them are in the developing world.

2. For more information on the IDRC (www.icpsr.umich.edu/IDRC/), see this press release, dated November 2007: www.icpsr.umich.edu/ICPSR/org/announce.html#idrc.

3. See www.internationalsurveynetwork.org/home/?lvl1=about&lvl2=members for a complete list of members. The Web site also has excellent sections on the legal issues on dissemination of microdata, the rationales for doing so, and suggestions for best practices.

4. Anonymization inevitably leads to some information loss, which is of course why students and researchers prefer to get at the confidential data! For more discussion, see www.internationalsurveynetwork.org/home/?lvl1=tools&lvl2=anonymization&lvl3=loss.

5. Taken from a typical form from the National Sample Survey Organisation (NSSO) from the government of India. See mospi.nic.in/mospi_nsso_undertaking_form.htm.