The impact of documentation on secondary data use

Jinfang Niu

Ph. D candidate

School of Information, University of Michigan

Introduction

Organizing, managing information, and providing information services to users have been long traditions in information schools. Numerous researches have been done on the organization and use of books, journals, images, even audio and video resources. Social science raw data, on which many publications and scientific findings are based, are important information resources. But they have received inadequate attention from information schools in the past. This is partly because the secondary use of raw data was not a common practice in many academic fields. -Facilitating data sharing has been a growing concern in recent years. In the United States, a law has been passed by the Congress to mandate the Office of Management and Budget to amend Circular A-110[1] to extend the Freedom of Information Act (FOIA) to "require Federal awarding agencies to ensure that all data produced under an award will be made available to the public under the FOIA"(http://www.whitehouse.gov/OMB/fedreg/a-110rev.html). More and more funding agencies require grantees to share their research data to the public. We can expect that in the near future, more and more social science raw

[1] Circular A-110 is a policy that implements FIOA.

data will be available for public access and use. Thus, there needs to be more research about the organization, management and use of social science data.

Secondary data use is defined this way: a user uses a data set, and the user is not involved in the production process of that data set. Three basic entities are involved in secondary data use: the data producers, data users, the data itself and associated documentation. Sometimes there is an intermediary between data producers and users. Intermediaries, such as data archives or data libraries, process the data to improve the quality of data and documentations, and disseminate data to users.

Documentation is metadata of social science data. Similar to MARC records that help users to search and judge the relevance of books and journals, Dublin Core records help users to search and determine the relevance of Web resources. Documentation provides information necessary to search and judge the relevance of data, more importantly, they help users understand and use the data. The Interuniversity Consortium for Political and Social Research (ICPSR) identifies the following as necessary elements of good documentations: cataloging information, such as title, principal investigator, data producer, place and date of production, funding agency; description of how the data were collected and the data sources used; full description of sampling design, frame, and methods as well as sampling error; full variable and value labels, full details of all coding classifications; question text, full description of recoded and derived variables, frequencies of variables, fully documented weights with information on conditions under which they should be used; and

details on file types and linkages among files. If available, data collection instruments and related bibliographies should also be provided.

Research Questions

There is a lack of empirical studies on the secondary use of social science data. Existing literatures about social science data sharing were mostly written in the 80's and 90's. More recent studies on the subject seem to be lacking. Most of the available literature discusses the need for data sharing, the costs, benefits and risks of data sharing, obstacles to share data, the legal environment of sharing social science data, what intermediaries do to help data sharing, etc. Some papers explore use experiences, but these mainly consist of individual opinions and arguments based on personal stories or anecdotal evidences, or summaries of discussions at conferences. While most literature mention that documentation~~s~~ is~~are~~ often insufficient for secondary use, they did not explain how the insufficiency affects secondary data use, or how users deal with insufficiency.

This study shall fill this gap by doing a systematic and empirical study on the secondary use of social science data~~, by~~ focusing on the impact of insufficient documentation~~s~~. It breaks use experiences into three aspects. First, users choose which data set to use. Second, users get the data, try to use it, and then decide to give up or continue to use it. Third, how difficult users feel about their using experiences. The aim is to examine the effects of documentations on each of those three aspects. The research questions are framed as follows:

1. How do users choose which data to use?

2. What are the reasons why users give up using or continue to use a data set?
3. What factors make documentation more or less sufficient for use?
4. What factors make data use experiences more or less difficult?

Methodology

Three steps are involved in answering these questions. First, from existing studies on knowledge reuse, knowledge transfer and the secondary use of non-social science data, a list of factors for question 3 and 4 has been identified. Second, qualitative interviews shall be conducted. Based on the interviews, answers to research question 1 and 2 shall be explored. In addition, the preliminary list of factors shall be refined by adding and deleting factors, breaking one factor into several sub-factors, or adjusting the factors to make them applicable for the secondary use of social science data. For the factors that will be analyzed quantitatively, quantifiable measures shall be established. In the end, a survey will be conducted

Respondents of study shall be the end users of secondary social science data. For the interviews, the convenient sample and snowball sampling methods shall be employed. A preliminary step is to contact known secondary data users. Further respondents shall be contacted based on the names recommended by the initial group.

Respondents to the survey shall be drawn from the ~~download records~~user pool of

a data archive.

For the first two research questions, the units of analysis are the individual data users. For the third and fourth research questions, the units of analysis are the single data sets in individual use cases. For each interviewees or survey respondents, general questions about their secondary data use experiences shall be asked to generate the answers to research question 1 and 2. Then, they will be asked to talk about~~t about~~ their recent experiences in using a particular data set. Through this procedure, data on two kinds of units of analysis can be generated through one study.

The first two research questions will be answered qualitatively. Question 3 and 4 will be answered both qualitatively and quantitatively. For the quantitative analysis of question 3, the dependant variable is the sufficiency of documentations. Below is the list of independent variables identified so far:

- Source of data: where the data is obtained from;
- Type of data: is the data survey/interview data, experimental data, longitudinal data etc.
- Producers of data: is the data produced by individual researchers, private organizations, or government agencies.
- The time of data production: when the data was produced.
- Knowledge distance: how much are the users familiar with the data,

the topic of the data, the data collection and analysis methods.

- Collaboration: how many other people are the users collaborated in using the data.

For research question 4, the dependant variable is the level of difficulty, which is users'perception of how difficult it was using the data. The independent variables are: the sufficiency of documentations, all the independent variables for question 3, plus two additional variables, the popularity and complexity of the data. Popularity is measured by the number of publications based on the data. The complexity of the data set is measured by the number of sub data sets, variables and observations in the data set, the actual number of variables and observations used by the users.

Significance of the Study

This is an attempt to provide the first systematic study on the secondary use of social science data. The empirical findings will benefit two parties. First, it tells secondary users what kind of data are better documented and easier to use, it also tells them how they should prepare themselves to make their data using experience easier. Second, it will inform data libraries or data archives what kind of data need more processing and what they can do to help users. This study categorizes social science data and its users, identifies factors affecting using experiences, and describes relationships among factors. In this sense, it is a step towards a theory of secondary data use. Moreover, I applied knowledge

management theories into secondary data use. Findings of this study will contribute to the theories of knowledge reuse and knowledge management.