What Data Characteristics are Needed for Data Reuse in the Domain of Social Sciences in Korea?

Nayon Kim[1][0000-0003-1993-7322] JungWon Yoon[2][0000-0003-2579-7688] and EunKyung Chung[1][0000-0001-9896-8623]

[1] Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760 Korea
echung@ewha.ac.kr
[2] University of South Florida, Tampa, Florida USA

Abstract. With the benefits of data sharing and reuse, data reuse have been promoted in various domains. While there are practices and discussions regarding data sharing and reuse, we still have little knowledge on what characteristics of data impact decisions on data reuse. In this sense, we aim to explore data characteristics in the context of data reuse within the domain of social sciences in Korea. For the purpose of this study, we conducted in-depth interviews with twelve researchers in the field of social science in terms of six dimensions: data producer, country/language, data type/collection method, procedure, accessibility, size/currency. For the producer dimension, social scientists preferred data that have been produced by an institution rather than an individual researcher. In language used in the data sets, English were more favored because researchers preferred English than any other languages. In terms of data type, quantitative and survey data types are preferred. For the procedure of data, researchers preferred original raw data with plenty of metadata and demographic information for analysis. For accessibility, there was less preference for restricted data. Lastly, for size/currency, researchers showed a preference for big size data and current data. These preliminary findings can provide better understanding about data reuse and guide improved data reuse services.

Keywords: Data Reuse, Data Needs, Data Characteristics, Social Sciences, Korea.

1 Background and Introduction

While the research environment is changing in an intensive manner, data are increasingly recognized as a research product, rather than simply a by-product. Thus, the sharing and opening-up of data is activated by governments and research institutions as the core of scientific discovery and technological innovation. However, there is a limit to promoting the reuse of data only by expanding data supply and ensuring accessibility. Rather, it is necessary to further promote the sharing and reuse of data by investigating what kind of data the researcher wants and providing the needed data. In [1], they investigated the factors influencing the perceptions and experiences of data reuse behav-

iors in the domain of social sciences. By analyzing the interviews from thirteen researchers who used Inter Consortium for Political and Social Research(ICPSR) and Dataverse Network, the author identified 25 factors. The findings show that there were six theoretical dimensions of influencing factors on data reuse. Practical and economic motivation, which considers advantages such as the time, resources, and money saving with reusing the data, was merely one part of those multi-layered dimensions. Other dimensions were also found, such as potential disadvantages of reusing the data, the perceived reusability of data, the effort required to handle the data, the availability of technical and personnel support, and the receptiveness of their peers when reusing the data. On the other hand, [2] investigated the experiences of success and failure in the domains of health and social welfare, and identified the obstacles of re-using the data. Based on in-depth interviews with a total of 23 researchers, the findings showed that accessibility, ease of use, documentation of data, and supplementary materials were significant factors whether data reuse is successful or not. [3] and [4] examined the factors on data reuse success. They focused on the novice researchers in the domain of social sciences in the first study [3], while they extended to a more experienced researcher group in later study [4]. In order to understand data reuse among the novice researchers, they interviewed 22 researchers who had used the data from the ICPSR. The findings showed that detailed documentation on data collection and coding, or experienced advisors on how to use the data were highly significant factors. A later study [4] conducted a survey questionnaire and collected a total of 249 survey results. Based on the analysis of the survey, they indicated that data completeness, accessibility, easy to use, credibility, and the quality of documentation were substantial factors. On the other hand, [5] focused on the data reuse behaviors of qualitative data, rather than quantitative data in the field of social sciences. With in-depth interviews of a total of 40 researchers, the findings showed that the most significant factor on the reuse of qualitative data was personal relationships such as student-advisor, peer, and alumni. As the related studies demonstrate, there have been endeavors and efforts to understand the data reuse behaviors in terms of affecting factors, success/failure, the status of researcher, and type of data for reuse. In this context, this current study aims to examine the needs of data reuse of Korean researchers in the field of social science.

2 Methods

For this study, two phases of data collection were conducted. First, we collected the article list which re-used the data provided by Korea Social Science Data Archive (KOSSDA). KOSSA is a non-profit data archive for social sciences and was established in 1983 in Seoul, Korea. Of the 1,546 data reuse articles during the period of 2014 to 2018, a total of 626 scholarly articles, which were indexed in Korea Citation Index (KCI), were selected for analysis. In the second phase, we emailed a total of 304 corresponding authors of the articles, which were published in the period between 2016 and 2018, regarding participating in this study. Then, we selected 12 researchers among the authors who accepted to participate in this study. The demographic information of the participants is shown in Table 1.

Table 1. Demographic information of the participants.

	Age	Sex	Degree	Major	Position/Affiliation	Career yrs.
P1	50s	M	Ph.D.	Sociology	Professor/University	20
P2	40s	M	Ph.D.	Economics	Senior Researcher/Research Institute	17
P3	50s	M	Ph.D.	Public Health	Professor/University	15
P4	40s	F	Ph.D.	Sociology	Professor/University	14
P5	50s	M	Ph.D.	Politics	Professor/University	12
P6	50	M	Ph.D.	Politics	Principal/High School	12
P7	40s	M	Ph.D.	Business Mgmt.	Professor/University	10
P8	40s	M	Ph.D.	Sociology	Professor/University	9
P9	50s	M	Ph.D.	Economics	Senior Researcher/Research Institute	8
P10	30s	M	Ph.D.	Politics	Junior Research/Research Institute	8
P11	40s	M	Ph.D.	Sociology	Lecturer/University	5
P12	30s	F	Ph.D.	Social Welfare	Senior Researcher/Research Institute	5

3 Preliminary Findings

Based on the previous studies of [6], [7], [8], and [9], dimensions of data reuse needs were adopted and modified: data producer, country/language, data type/collection method, procedure, accessibility, and size/currency. The preliminary findings from the interviews with 12 researchers were analyzed.

3.1 Data Producer

Most of interviewees preferred data produced by institution, rather than individual researcher. More specifically, government, international organization, government research institute, and university research institute were found as preferable producers among researchers. The reasons for this preference of institute are primarily classified as accessibility, comparability, and credibility.

"*…it is because the data is accessible, in fact, I think, company data might be better, but it is hard to get, but the data from this institute can be accessible all the time…*" (P2, researcher, Economics)

"*…I think it is possible to compare multiple countries or cultures if I use this kind of data…*" (P8, Professor, Sociology)

"*…I can trust the quality of data produced by this institute, because they have produced this data for a long time. The data collection is systematic, so the quality seems to better than other data sets…*" (P12, researcher, Social Welfare)

3.2 Data Country/Language

Researchers are likely to reuse the data from the countries such as US, Japan, China, and UK. In addition, the countries from EU and member countries from OECD are preferred. One of primary reasons for this preference is that researchers tend to publish their results in the journals in those preferable countries.

"*…I would choose the data produced in the United States because I plan to submit the manuscript to one of journals from the United States…*" (P11, Lecturer, Sociology)

On the other hand, researchers tend to prefer the data in English if the data is not in Korean. The main cause for this choice is their language capabilities.

"*…previously, I had hard time to use the data in Japanese, because it was very hard to use the data in Japanese. If the data was in English, I believe that I would use more efficiently, because I feel confident with using the data in English…* " (P6, Principal, Politics)

3.3 Data Type/Collection Method

It was found that data reusers strongly preferred quantitative data, rather than qualitative data. The preferred data collection methods for quantitative data sets were identified as personal survey and ARS phone, rather than online web survey, because of the quality of data issue

"*…when data collection was conducted, depending on who collect the survey answers, there were big differences in terms of the quality of survey answers, so the quality of data…* " (P10, Researcher, Social Welfare)

3.4 Data Procedure

The data procedure includes the level of aggregation, the metadata, and identification information. In terms of aggregation of data, researchers tend to prefer the raw data, rather than sample data and analyzed data.

"*…there is no reason to analyze the sample data because computing capabilities are powerful nowadays…* " (P1, Professor, Sociology)

For metadata, researchers show a strong preference on the data sets with rich metadata. This finding is consistent with the result of [10].

"*...I believe that the significance and credibility of data were associated with the quality of data. They (metadata) are very important when we use the data set ...*" (P8, Professor, Sociology)

Researchers are likely to use the data with demographic information of the participants such as age, sex, occupation, education, income, marriage, and location. However, researchers noted that lots of data were provided without the location information of the participants because of the privacy law in Korea.

"*...for instance, the data were collected in the level of Dong, [which is the smallest administrative unit in Korea], but the unit in the released data was deleted. If the unit information was provided, then the data could be analyzed with some level of context ...*" (P8, Professor, Sociology)

3.5 Data Accessibility

Researchers are not likely to overcome the barriers of accessibility when the data are under restricted access or require special permission.

"*...there was a case that if I asked the author of the data, then the author advised me to ask the institute of the data, if I asked the institute of the data, then they said that I need to ask the author of the data, which is very inconvenient for me...*" (P6, Principal, Politics)

3.6 Data Size/Currency

Researchers show preference for the big data sets and current data. When the statistical analysis is involved, the size of data does matter to the researchers.

"*...the more, the better. I prefer the big size data such as labor panel data and company data, which are between 15,000 and 20,000 participants and 4,000 participants, respectively...*" (P6, Principal, Politics)

For the currency of data, researchers tend to use the most current data because the result research paper should reflect the current status of the social phenomenon.

"*...I think that social sciences change fast, so the data should be current, otherwise it would be useless for understanding the society ...*" (P6, Principal, Politics)

4 Discussion and Conclusion

In the data-intensive current academic setting, data has become a foundation for scholarly communications as significant research result. However, the mere expansion of data supply, the availability, and the accessibility cannot always lead to practical data reuse. Moreover, due to the data deluge, the data management and utilization ability of individual researchers or research institutes has failed to follow their production capacity. Therefore, the needs for data reuse should be understood in various perspectives. In this sense, this study aimed to explore the characteristics and values of data reuse in academic research activities and examine the data needs of academic researchers and identify the data attributes preferred for data reuse. With preliminary analysis, we identified six major factors on data reuse. First, institutional producers, rather than individuals, were preferred for suitability, comparability and reliability of the data. Second, priority was given to data produced in the United States or written in English, as this would ensure comprehension of the data. Third, for the quality of data such as representativeness and reliability, priority was given to survey data, especially data produced through investigator-filled survey, rather than online survey data. Fourth, efforts should be made in several ways to obtain a wealth of metadata. In addition, priority was given to data that contained identification information at a level where demographic and sociological characteristics could be understood within legal limits. Fifth, access and use-controlled data were not likely to be sought for reuse. For the sixth, more preferences were on big, recent datasets and, rather than small samples and old data, especially when statistical analysis was involved. As the preliminary findings demonstrate, there were distinctive characteristics of data in terms of data reuse. Since the availability of data does not directly link to the usage of data, the characteristics of data should be utilized when data reuse services are provided. Further analyses on the interviews will be conducted for better understandings on the needs and behaviors of data reuse.

References

1. Curty, R. G.: Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study. IJDC 11(1), 96-117. (2016).
2. Yoon, A.: Red flags in data: Learning from failed data reuse experiences. In Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology (p. 126). American Society for Information Science. (2016).
3. Faniel, I. M., Kriesberg, A., & Yakel, E.: Data reuse and sensemaking among novice social scientists. Proceedings of the American Society for Information Science and Technology 49(1), 1-10. (2012).
4. Faniel, I. M., Kriesberg, A., & Yakel, E.: Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology 67(6), 1404-1416. (2016).
5. Yoon, A.: Making a square fit into a circle': Researchers' experiences reusing qualitative data. Proceedings of the American Society for Information Science and Technology 51(1), 1-4. (2014).
6. Nicholas, D.: Assessing information needs: tools, techniques and concepts for the internet age, Aslib Information Management. (2000).

7. Borgman, C.L.: Big data, little data, no data: scholarship in the networked world. MIT press. (2015)
8. Zimmerman, A.: Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries 7(1-2), 5-16. (2007)
9. Niu, J.: Overcoming inadequate documentation. Proceedings of the American Society for Information Science and Technology 46(1), 1-14. (2009).
10. Niu, J., & Hedstrom, M.: Documentation evaluation model for social science data. Proceedings of the American society for information science and technology 45(1), 11-11. (2008).