Using Open Data for Social Sciences

Akio Yoshida [0000-0003-1001-314X]

Jawaharlal Nehru University, New Delhi, Delhi 110067, India

akio.yoshida@gmail.com

Abstract. Information and communication technologies (ICT) are changing research methods in social sciences, especially in the ways of getting data. Online surveys and Big Data analyses are used the most among them. Using Open Data is another way realized by the spread of ICT. Using Open Data by administrative agencies has potential but also some difficulties. This study discusses the practical use of Open Data, and focuses on problems related to data in Portable Document Format (PDF) files. Those problems seem to occur because many officers in charge of Open Data do not pay attention to the principle and practical use of Open Data. It shows a gap between drafters and practitioners in our society.

Keywords: Research Method, Social Sciences, Questionnaire Survey, Big Data, Open Data, File Format, PDF.

1 Introduction

Social data can be used to improve and help researches in the field of social science. Those data exist in a variety of shapes. They are texts, voices, photos, movies, etc. They are also called materials, records, sources, and evidences. However, analyzing data in so many different formats presents a challenge. Therefore, one approach would be to try to quantify the collected data. With numeric data, results from objective views can be shown in statistical ways.

The rise of ICT brought changes to the research methods in social sciences. In addition to online questionnaire surveys and Big Data analyses, Open Data analyses are used recently. Contrasting with former two, this study shows the potential of using Open Data, as well as the practical problems of using them.

1.1 Questionnaire survey

If questionnaire was made well with options or Likert scales, it is possible to collect numeric data, as the results of such surveys are categorical and quantifiable. In the past social sciences have often relied on physical or offline questionnaire surveys.

Nowadays, online surveys with questionnaire are also a popular way for gathering user data. Respondents can answer to the questionnaire with smart

phones or tablet terminators. However, there have been problems in online surveys compared with offline surveys. Typical cases are summarized in Table 1. The online survey respondents are automatically limited to the Internet users who have registered to a survey company. We cannot reach those who do not use the Internet and those who are not registered to any Internet services. In addition, online direct mails are easy to be ignored. It leads to low recovery rate. Therefore, online surveys have been regarded being biased in sampling.

Table 1. Online vs Offline in typical social surveys.

	Online	**Offline**
Object	Internet user	General people
Sampling	Registered participants	Random
Delivery	Online DM	Postal service
Recovery	Reply form	Visiting
Recovery rate	Low	High
Sampling Bias	High	Low
Costs	Low	High

On the other hand, conventional offline surveys also got problems recently. Mail survey in Table 1 has been an efficient data collection tool since 1788 [1]. These days nuclear families, which consist only of parents and children, are increasing especially in urban area. When researchers visit their houses in daytime, they may not be able to recover questionnaires because no family members are available there. It is also difficult to reach young people who live alone. In addition, people are unlikely to open the door to unknown person's visit. These problems lead to sampling bias. Respondents come to be limited to those who can react to researchers in such circumstances.

When almost all the people come to use the Internet, online survey may have less sampling bias than conventional survey. Visiting respondents in offline survey can secure reliable response, while online submission can reduce Hawthorne effect that respondents tend to give desirable answers, which let them look more normative [2]. Each of these methods has its pros and cons.

1.2 Big Data

With improvement of ICT and computing, we came to treat Big Data, that is, huge amount of transactions of information. Including GAFA (Google, Amazon, Facebook, Apple), many large corporations utilize Big Data which they collect in their businesses [3]. Using Big Data has a tremendous potential to benefit social sciences. However, those studies of Big Data are in a black box format. Some

companies apply the findings to their businesses; others optimize them as B2 2B commodities. Analyses of data and the findings are not disclosed. Though SNS companies such as Twitter provide API to get data, there are limitations of using them.

Companies are eager to protect their algorithms, but at the same time, the data itself can contain sensitive information, for instance transaction data or personal customer information. That is why many countries are developing legal systems on personal information (Table 2). Japan amended the Act on the Protection of Personal Information for use of Big Data. It includes "Clear Indication of the Purpose of Use", "Consent of the Person on Provision to A Third Party" and "Anonymization of Information". General Data Protection Regulation (GDPR) in EU has more rules. It is said that DPA in Kenya and PDP in India are based on GDPR [4][5].

Table 2. Legal systems on personal information.

	Laws	**Enforcement**
USA	Sector and State Specific Rules / FTC: Federal Trade Commission Act § 5	- / 1914 *2012
Japan	Act on Protection of Personal Information	2003 *2017
China	CS: Cybersecurity Law	2017
EU	GDPR: General Data Protection Regulation	2018
Kenya	DPA: The Data Protection Act	2019
India	PDP: Personal Data Protection Bill 2019	Pending

*amended

In June 2013, Hitachi announced that it would start a service that utilizes the boarding / alighting history of JR East (Japan's largest railway company)'s Suica (IC prepaid fare card) as big data and provides it as station area marketing information. At first, JR East claimed that it was not disclosing personal information on its customers, but admitted selling data without their consent and apologized after a month [6]. It is still withheld in 2021.

In January 2021, messaging app WhatsApp announced the new Privacy Policy, which will allow WhatsApp to share data with its parent company, Facebook. It does not apply in EU, because it violates GDPR [7]. People encouraged each other to shift from WhatsApp to other messaging apps, Signal or Telegram. At last, WhatsApp postponed the update of its privacy policy.

In March 2021, LINE, which is very similar to WhatsApp and dominant in Japan, let Chinese engineers at a Shanghai affiliate access Japanese users' data without informing them [8]. LINE Corporation was founded as a part of a South Korean game company.

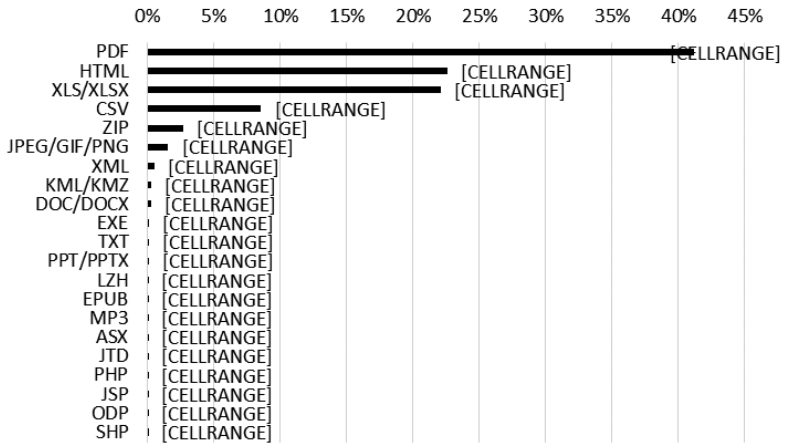
Big Data can be international. It is important to pay attention to the latest trend in the world. Even if the use of data is legitimate in Japan, it may violate GDPR in EU [9]. Though these legal systems encourage the use of Big Data, companies will be careful with further use of them. Moreover, it will take time for a broad range of academic use of Big Data.

2 Using Open Data

We have already covered some of the challenges when it comes to collecting numeric and large amount of data in social sciences. Using Open Data could present an alternative way of data collection. Knowledge is open if anyone is free to access, use, modify, and share it [10]. Usually there are limits and difficulties to access data, as mentioned in previous section. Even after getting data, there are still problems as license, copyright, patent or other mechanisms of control. Open Data are free from these restrictions. There are two kinds of data expected to be Open Data. First, they are academic data in sciences. Second, they are social data obtained by administrative agencies.

Using academic Open Data is, in other words, the secondary use of data. The data of GSS (General Social Survey) in USA are generally available in formats designed for statistical programs, and "GSS Data Explorer" allows users to test hypotheses, and look for interesting correlations directly on the website.

Social data obtained by administrative agencies are also published and free to access by the public. According to a questionnaire survey in Japan, medians of Open Data government possession rates were only 1% to 5% in each section: spatial Information, Agroforestry, Commerce and Industry, Medical and Welfare, Education Tourism, and Others [11]. This means Open Data by local governments have a big potential. Most of data by governments are census data. They are free from sampling bias in social survey. When they are published, that certifies they are free from the problems of private information in Big Data by private companies.



Source: DATA.GO.JP (*on Mar 26, 2021*)

Fig. 1. Numbers of file formats in Open Data by the central government of Japan.

By way of illustration for a problem in Open Data by administrative agency, there have been arguments on official announcement about the results of national academic ability survey in Japan. When the governor of Osaka prefecture, Hashimoto decided to publish the data by cities, towns and villages, some municipalities and activists were against it. When he became the mayor of Osaka city, he disclosed the results of the city by schools. Now results in 2011 and 2012 are available except municipalities with only one school [12]. An academic use of the data considered not to identify those schools [13].

As mentioned at Introduction, Open Data by administrative agencies are not only numeric but can also come in the form of documents. They are provided in PDF (portable document format) files. There were 9776 PDF data sets in Japanese data catalogue site "DATA.GO.JP". They enabled cross-sectional search of the data by the central government [14]. That made up about 40% of all the data sets in the site. After 4 years, while the data sets increased by 50%, the rate of PDF format keeps still 41.2% of all (Fig. 1).

These PDF files often are not machine-readable, even if they have literal or numeric data. When we retrieve them, we may need OCR (optical character recognition) software. For example, Election Commission of India has data of donation, which have tables with donors and amounts. However, they are not machine-readable. After retrieving data with software, we have to review the error rate of the OCR algorithm, with viewing operation. They are scanned data from paper documents, which were printed out. Punch holes in sequential documents often damage some parts of data. The spread of paper-less transactions in administrative agencies may solve these problems.

Table 3. Star scheme* toward Linked Open Data by Berners-Lee.

★	Available on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

*added in 2010

Source: Linked Data [15]

Berners-Lee, known as the inventor of the world wide web, developed star rating system "*in order to encourage people -- especially government data owners -- along the road to good linked data*" [15] (Table 3). This table present a scale, well known to officers in charge of Open Data in governments.

PDF format is supposed to be worth 3 stars in a manner independent of application software, hardware, and operating systems. However, many data in PDF will not get even 2 stars because they are not machine-readable. In the survey on local governments in Japan [11], it was discussed whether machine-readable PDF should be distinguished from non-machine-readable one in the questionnaire. Officers in charge of a certain municipality said, "They may not be aware of the difference between normal PDF and image PDF. PDF may be only PDF for them." It was considered that such a question could be difficult to answer by respondents – if they do not have the necessary background to make this distinction.

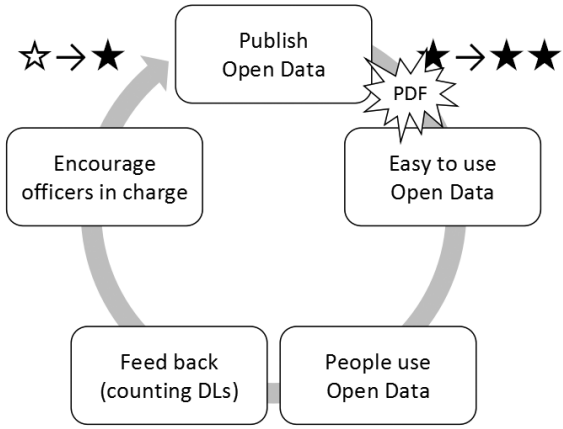


Fig. 2. A Cycle toward more Open Data related with Star scheme and the PDF problem.

Though Berners-Lee drew a blueprint of Linked Open Data, the star scheme he developed does not cover all current problems connected to open data. For instance, to get Linked Open Data, we need more Open Data. The value of data will increase, if there are more related data [16]. However, those data are not easy to use, unless they are machine-readable. If they are not easy to use, people may not use them. There is a structured interview research, which showed data users could motivate officers in charge of Open Data [17]. Emotion of officers cannot be overlooked. The above-mentioned officers also said, "It is pleasure for public servants that people use Open Data. It encourages us to contribute for public interest." They know how many times their Open Data sets were downloaded. If people use more data, officers may publish more data (Fig. 2). The difference between 1 star and 2 stars is very important as well as that between 0 star and 1 star. Here are proposals to Star scheme.

- There should be an instruction to distinguish non-machine-readable PDF from machine-readable PDF.
- Machine-readable should be translated to "possible to copy and paste textual data or matrix data" for ordinary people.
- PDF should be included as examples, as well as CSV and excel.
- "2 stars system" can highlight the importance of the difference between 1 star and 2 star. It can be more efficient to encourage officers in charge toward Linked Open Data, so far.
- Open Data providers can share 3 to 5 stars works to the third parties or Open Data catalogue site.

3 Conclusion

This study highlights some of the many challenges involved in collecting numeric data in social sciences, the actual conditions of Open Data by administrative agencies, and a practical use of Open Data. There are changes in research methods with ICT. Conducting a social survey is getting difficult. Utilizing Big Data for academic purpose still presents many problems, which need to be solved. At the same time, using Open Data by administrative agencies has a tremendous potential. Open data is a source for a high volume of free documents. Many document data are made by scanning printed documents. They are published in non-machine-readable PDF files. People may not pick up such Open Data, which are hard to use. More use of Open Data can generate more Open Data from public sectors. Therefore, some proposals on the problem of Open Data in PDF files were presented. Though the problem of a file format in this study looks very trivial, it may have prevented the spread of Open Data. The Star scheme was made to encourage officers in charge of Open Data. It has been well known to them, its principle still does not seem to be realized by them even after a decade.

It must be significant to have pointed out a gap between drafters and practitioners in our society.

This study only pointed out the existence of the problem of PDF. It was discussed only with cases in Japan and India. It was not examined whether the problem exists all over the world, and how many non-machine-readable PDF there are. There can be some reasons that officers in charge tend to make image PDF files. For example, they may be going to put priority on signatures or stamps. Convenience is not always right. It should be discussed with Electronic Signature together. These points should be improved and will be the future works.

References

1. de Heer W., de Leeuw E.D., van der Zouwen J.: Methodological Issues in Survey Research: a Historical Review. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 64(1), 25–48 (1999).
2. Landsberger H. A.: Hawthorne revisited: a plea for an open city. Ithaca, N.Y.: Cornell University (1957).
3. Marr B.: Big Data in Practice: How 45 successful companies used Big Data Analytics to Deliver Extraordinary Results. Chichester, Wiley (2016).
4. Kazeem Y.: Kenya is stepping up its citizens' digital security with a new EU-inspired data protection law. Quartz Africa, November 12 (2019) https://qz.com/africa/1746202/kenya-has-passed-new-data-protection-laws-in-compliance-with-gdpr/, last accessed 2021/03/30
5. Jain R.: An existentialist dilemma for the Non-Personal Data regulation?," Telecom.com, March 23 (2021). https://telecom.economictimes.indiatimes.com/tele-talk/an-existentialist-dilemma-for-the-non-personal-data-regulation/4861, last accessed 2021/03/30
6. Metcalfe J.: Japan Railway Company Apologizes for Selling IC Card Data. The Wall Street Journal, July 29 (2013). https://www.wsj.com/articles/BL-JRTB-14515, last accessed 2021/03/30
7. Lakshmanan R.: WhatsApp Will Disable Your Account If You Don't Agree Sharing Data With Facebook. The Hacker News, Jan 6, (2021). https://thehackernews.com/2021/01/whatsapp-will-delete-your-account-if.html, last accessed 2021/03/30
8. Reuters: Japan to probe Line after reports it let Chinese engineers access user data. March 17 (2021). https://www.reuters.com/article/us-japan-line-access-idUSKBN2B901E, last accessed 2021/03/30
9. Terada S.: Overview of foreign legal systems related to personal information protection. JIPDEC (2019). https://www.jipdec.or.jp/archives/publications/J0005156.pdf, last accessed 2021/03/30
10. Open knowledge Foundation: Open Definition 2.1 https://opendefinition.org/od/2.1/en/ last accessed 2021/05/15
11. Noda T., Honda M., Yoshida A.: Economic Effect by Open Data in Local Government in Japan," In: Baghdadi, Y. and Harfouche, A. (eds.) ICT for a Better Life and a Better World, The Impact of Information and Communication Technologies on Organizations and Society. pp. 165–173. Springer, Heidelberg, (2019).
12. Osaka prefecture.: Public elementary school, junior high school and kindergarten. http://www.pref.osaka.lg.jp/life/list2.php?ctg02_id=18, last accessed 2021/03/30
13. Uesugi M., Yano K.: A Geodemographic Analysis to Assess Variations in School Performance Based on Educational Achievement: A Case Study of Osaka City, Japan. Japanese Journal of Human Geography (Jimbun Chiri), 70(2), 253–271 (2018)

14. Honda M.: The whole aspect of public data to suppose from “DATA.GO.JP”. Journal of Japan Society of Information and Knowledge, 26(4), 320–325, (2017)
15. Berners-Lee, T.: Linked Data. (2006). https://www.w3.org/DesignIssues/LinkedData, last accessed 2021/03/30
16. Shapiro C.: Information rules : a strategic guide to the network economy. Varian, Hal R. Boston, Mass. Harvard Business School Press (1999)
17. Honda M., Kajikawa Y.: Importance of communication between policy makers and external actors in the policy formation process. Proceedings of the 15th National convention of Japanese Association for Communication, Information and Society. pp. 204-207 (2018)