Overcoming Inadequate Documentation

Authors

Jinfang Niu
University of Michigan
800 Blaine Ct. Apt# 1702, Schaumburg, IL 60173
Email: niujf@umich.edu

Secondary data users need three types of knowledge to analyze secondary data: knowledge about data, background knowledge necessary to understand and interpret data, and data analysis skills. Part of knowledge about data is provided by the documentation of data. Background knowledge and data analysis skills are internalized as users' absorptive capacity. When documentation and their absorptive capacity are inadequate, users need to seek outside information to use secondary data. In this paper, causes of inadequate documentation were analyzed, why and how secondary users seek outside information were reported. Then based on the findings, implications about how to facilitate secondary data use were discussed.

Introduction

Secondary data use means analyzing data produced by other people. To enable secondary data use, data and knowledge about data need to be transferred from data producers to secondary users. Documentation plays an important role in transferring knowledge from data producers to secondary users. Previous research (Niu and Hedstrom, 2009) has found that perceived documentation quality varies with several characteristics of data and is weakly affected by users' absorptive capacity. In this paper, I will focus on how users overcome inadequate documentation by seeking information beyond what is provided with documentation. I will also provide some recommendations about how to facilitate secondary data use. Findings reported in this study are based on data collected for a larger research project. Details about the data collection method were reported in (Niu and Hedstrom, 2009). Here I briefly review it: I interviewed 13 secondary data users for 30 to 70 minutes each. The interviews were unstructured and exploratory. Secondary users were asked questions about their secondary data analysis experiences in general and their most recent experience with a particular dataset. Preliminary findings from the interviews informed the next step survey design. The formal survey started in May 2008. Each

respondent was asked about his or her most recent experience using a dataset produced by a different individual or entity. 1,260 surveys were sent out and 384 usable responses were received. The survey collected both highly structured quantitative data and open-ended qualitative data.

Knowledge1 Necessary For Secondary Data Use

Based on the kinds of information that users seek during secondary data use, I identified three types of knowledge necessary to analyze secondary data: knowledge about the data, background knowledge used to understand and interpret data, and data analysis skills. Below are some examples of the three types of knowledge.

- Knowledge about data: what is the response rate and sampling frame for a particular survey? how are the missing responses treated?
- Background knowledge: disciplinary consensus on how to use common types of data, how to determine whether or not to weigh variables from samples? which variable best captures certain concepts? how to interpret frequently occurring variables? how to handle specific measurement issues?
- Data analysis skills: how to convert hierarchical data files into appropriate rectangular files? or how to construct new derived variables?

Part of knowledge about data is provided by documentation of data. The term "documentation" is used for knowledge about data recorded and transferred to secondary users that helps secondary users understand and use data. Here are exemplary documents that can be parts of documentation: codebooks (sometimes called data dictionaries), reports about the data collection project, data collection instruments, previous publications based on the data, user guides or handbooks, statistical manuals, data extraction software, programs making new variables based on the original data, original IRB materials, workflow for creating new datasets based on existing data, etc. Documentation plays an important role even when people involved in producing and documenting data are present in secondary data use. interviewee #1 struggled with inadequate documentation even though she was collaborating with the Principal Investigator of that dataset, because the research assistant who did most of the data collecting and documenting work had left the project. Interviewee #7 was part of the data analysis team for the National Survey of Child and Adolescent Well-being. She was involved in writing project reports for that dataset. However, in her later research based on that dataset, she still needed to go back to the reports that she wrote for more information.

Background knowledge and data analysis skills are accumulated through academic training and professional experiences. Unlike documentation, which is external knowledge, background knowledge and data analysis skills are internalized as part of secondary users' knowledgebase. I borrow the term "absorptive capacity" from Szulanski (1996) to refer to knowledge necessary to use secondary data and internalized by secondary users. Knowledge from documentation can be internalized and become a user's background knowledge. In those cases, users would be less reliant on documentation. This is most dramatic when a user has used the same dataset or the same data series many times.

Why Do Users Seek Outside Information?

The previous section tells us users rely on both external documentation and their internal absorptive capacity to analyze secondary data. Consequently, the inadequacy of either documentation or absorptive capacity could motivate users to seek outside information to fulfill their needs. In this paper, I use the term "outside information" to refer to knowledge and information that is necessary for secondary use but is missing from documentation or users' absorptive capacity.

The adequacy of documentation can be measured on three dimensions: sufficiency, ease-of-use and accuracy. Documentation can be adequate enough so that users with sufficient absorptive capacity can use data solely based on documentation. Actually 19% of survey respondents were able to use secondary data based only on documentation. However, seeking outside information is often necessary. About half (46%) of surveyed users obtained outside information because the documentation did not contain the information they needed (not sufficient). Besides, hard-to-use documentation also turns users to other information sources. As one user said: "the code book contains the definitions of variables, but sometimes I think it is easier to pick up the phone. I got the codebook, but sometimes the variables are not clear to me. They do have a user manual, as well as statistical manual. Sometimes if I know someone who might know the answer, it is easier for me to pick up the phone." 12.5% of surveyed users sought outside information because the documentation was hard to use. 31.4% of the survey respondents obtained outside information because other information sources or channels were immediately accessible, which is useful for understanding the following phenomenon: even though data produced for sharing are better documented than data produced for self-use (Niu and Hedstrom, 2009), users of data produced for sharing are not less likely to seek outside information than users of data produced for self-use. The reasons could be: producers who produce data for sharing tend to provide user assistance and make outside information more readily accessible. A few users sought outside information because they detected some errors in data and documentation (accuracy issues).

People who did not seek outside information rate documentation as more sufficient (z=5.7, p=0.00)[2] , easier to use (z=4.2, p=0.00) and of better overall quality (z=2.6, p=0.01) than users who did.

Some information seeking is driven by absorptive capacity. Users with lower professional status[5] are more likely to seek outside information (z=2.5, p=0.01). This is consistent with interview findings: students often get data analysis help from their advisors. As mentioned before, absorptive capacity is accumulated through academic training and professional experience. Therefore most of it is achieved before users start up using a particular dataset. As Interviewee #5 said: she would not be at the point of getting and analyzing data if she is not familiar with some terminology in the data. But often the case is that users did not expect something in a dataset during the process of analyzing it. For example, interviewee #6 used court records. Even though her field is criminal justice, sentencing and policing, she still needs to learn something about court process to use the data. To analyze a dataset using a new statistical method, interviewee #7 took a summer course offered by a data archive. Interviewee #5 received a lot of help from her advisor about how to read and interpret the data, and to select variables and conduct analysis while she analyze secondary data the first time. Here are more examples from the survey: "I did not receive information (about the data) outside of the documentation. What I received was additional information on how others used/interpreted the same data. This provided me with a deeper understanding of the data which I think was a benefit when I used the same data sets." "The information was not on the datasets itself but rather on operation of derived variables in previous studies. For instance, how do you handle negative incomes and the like." "It was an analytic question more than a data-specific question."

Besides the inadequacy of documentation and absorptive capacity, there are some social psychological reasons that users seek outside knowledge. Look at the following reasons that users sought outside information. "Working with others helps to bring the data alive." "Just want to get as much information as possible" "It was nice to talk to someone about the data instead of just read about it." "It is useful to talk with other secondary data users who are more knowledgeable about the data." "Website provided additional information that was useful for my analysis but I didn't need to use the website to use the data." There are also some users (11.6%) who obtained outside information not because they actively sought it, but because they happened to encounter that information.

Causes of Inadequate Documentation

Inadequate documentation is caused by two reasons. First, data are poorly documented because

data producers are not willing to or are not capable of providing adequate documentation for secondary use. Second, documentation is inherently inadequate due to the nature of tacit knowledge and communication reduction.

Researchers who produce data for self-use[6] often are not willing to spend effort to document data for secondary use. Producers" incentives to share and document data depend on whether mutually beneficial or mandatory sharing mechanism exists. Simply speaking, A would be willing to share with B if B will reciprocate in a certain way. A would have to share with B if there is some pressure forcing him to do so, which could be cultural pressure from community norms, data sharing policies, or grant conditions. For example, a group of researchers used game players" data from Sony Online Entertainment to study the dynamics of network and group behaviors. When asked why Sony was willing to share the data with them, they answered because Sony expected to benefit from their research. One of our interviewees shared his data because his funding agency requires him to share, and he was afraid that if he did not share, his chances of getting future grants would be damaged. Without a mutually beneficial mechanism, even if mandatory sharing policy exists, data producers can "comply with the letter of the law rather than its spirit, depositing poorly documented data that of little value" (Borgman, 2007, p. 242).

Resource constraint is a factor that keeps data producers from creating good documentation. Most respondents of our 2006 survey[7] expected that more time and financial support from funding agencies could be provided for documenting data. Some data producers complained that their grant was even insufficient to cover research needs, let alone creating documentation. By the time that the data are complete and the reports are delivered, time and funding are usually used up, with nothing left for documenting and sharing data. In addition, it is more difficult to document data for other people than for self-use, and that also challenges data producers' ability to provide adequate documentation for secondary use. According to Borgman (2007, p. 167), the effort required to explain one's data adequately increases as a function of the knowledge distance between data producers and users. Documenting research data for use by team members is more difficult than documenting it for personal use. Documenting it for off-site collaborators is more difficult still. Most difficult of all, however, is documenting for unknown future users, which is precisely the case for public data sharing.

Another reason is the nature of tacit knowledge and communication reduction. Existing literature suggested the following types of tacit knowledge. (1) Knowledge that is technically difficult to articulate. Tacit knowledge is hard to formalize and communicate because it is deeply dwells in a comprehensive cognizance of the human mind and body (Polanyi, 1962). (2) Tacit knowledge that is sensitive and subtle, even though people may know implicitly, it is not appropriate to make

them explicit. In talking about why groupware fails, Grudin (1993) said a priority-based meeting scheduler foundered because participants were reluctant to acknowledge publicly that some of their meetings were low priority. (3) Knowledge filtered out through communication reduction. Not everything can, or should be transferred. Some kind of reduction, and thus loss of complexity is inevitable (Strathern, 2005; Carlson and Anderson, 2007). Documenting data is necessarily a filtering process that only keeps the details that matter most to data producers. (4) Knowledge taken for granted by data producers. Data producers may unconsciously keep the details of their data collection and variable construction processes and the particular quirks of the data in their memories and do not put them in writing, without realizing that secondary users do not know those details (Fienberg, Martin, Straf, 1985). My survey and interviews revealed other categories of knowledge that tends to be missing from documentation. (5) Informal knowledge. One user said: "People don't document why some of the numbers are funny, things that went wrong in the survey, etc." (6) Missing knowledge caused by mismatches between the concerns of data producers and that of data users. As one user said: "while you are using existing data, most of the time somebody collected it for a different reason. The failure that I had with the sentencing data was that whoever collected that data, for whatever reason, didn't need to know where the offenders came from. So they didn't record it."

Where Do Users Seek Outside Information?

I categorize knowledge transfer channels into three types based on the kinds of knowledge that can be transferred through them. One channel is the use of documents. Only explicit knowledge can be transferred in this manner. A second channel is interactive conversations, such as face-to-face or phone conversations, meetings, or email messages. Through this channel, a receiver might be able to capture some tacit knowledge through the facial expressions or tones of the sender, but knowledge transferred through this channel is primarily explicit knowledge that is verbalized and not formally documented. When documents are not sufficient to transfer knowledge, conversations may help the receiver to obtain more information or further clarification. A third channel is situated learning. A typical example is an apprentice working with his/her mentor in order to learn craftsmanship not only through language, but more importantly, by observation, imitation, and practice. Tacit knowledge that is technically hard to articulate or socially sensitive can be transferred through this channel. This rationale is used in analyzing the sources and channels where users obtained outside information.

Table 1 is a list of information sources and the percentage of users who used each source for outside information. Please be informed that one user may seek information from several sources.

Table 1. Sources for outside information (N=353)[9]

Previously written articles using the dataset	47%	Documents (64%)
Websites of data producers	34%	
Websites of data archives	17%	
Data producers	41%	People (68%)
Other secondary users	40%	
Data archivists	13%	
Workshop	8%	

Besides those main sources, other information sources include: using related datasets to check the integrity of the data used or for other reasons, publications based on similar data collected by different researchers, outside sources of scales used in the data set, alternative publications with similar information, newsletters mailing lists for users of the same data, relevant newspapers, etc. 64% of survey respondents obtained outside information from various kinds of documents. Based on our rationale, that kind of knowledge is explicit knowledge that can be incorporated into documentation, or at least pointed to from documentation.

68% of survey respondents sought outside information from other people. Some of those people work closely with secondary data users, such as mentors, advisors and colleagues. Some are strangers, such as other users of the same data on a mailing list, other data users or data producers found through Internet search, data producers who left contact information in documentation, data archivists where users obtained data. Users who obtained data from data producers are more likely to seek outside information from data producers ($chi2=6.87$, $p=0.01$) and websites of data producers ($chi2=5.59$, $p=0.02$). Among 239 users who obtained information from people, 80% obtained that information through email or telephone, 55% obtained that information through face-to-face conversations, 31.4% obtained that information by working together with other people. 18.4% of users used all of the three channels. About half of the users (49%) used at least two channels. 36% only used email or telephone. 11% only used face-to-face conversations. 3% obtained knowledge only by working together with other people.

For users who obtained outside information only through one channel, Table 2 shows the distribution of reasons why users sought outside information. The five letters represent five reasons why users obtained outside information. A and B are related to the adequacy of documentation. C and D are related to the convenience of information sources and channels. We can see several patterns from the table. First, people are more likely to obtain outside information through email and telephone when documentation is inadequate. Second, people are more likely to obtain outside information through face-to-face or working together when those channels are convenient. Third, people are more likely to obtain hard-to-document tacit knowledge through working together. This third pattern is consistent with our rationale that tacit knowledge is more suitable to be transferred through situated learning than through documents or interactive conversations. Among 21 users who sought outside information only because of tacit knowledge problems, 18 sought that information from people (17 through email and telephone, 10 through face-to-face conversations, and 3 through working together); 14 obtained outside information from various documents (websites and publications based on the datasets); 7 only obtained information from people, 3 obtained that information only from documents.

Table 2. Reasons for obtaining outside information (N=353)

	Email/ telephone		Face-to-face		Work together	
A: Documentation does not contain information they need.	66%	72%[1]	31%	35%	29%	29%
B: Documentation is hard to use.	16%		8%		0%	
C: Other information sources and channels are immediately accessible.	28%	31%	58%	65%	57%	71%
D: They happen to encounter that information.	5%		19%		29%	
E: That information is tacit knowledge that is hard to document.[2].	18%		19%		43%	

A higher percentage of qualitative data users reported tacit knowledge problem than quantitative data users (chi^2=3.3, p=0.07). The fact that hard-to-document tacit knowledge is transferred also through documents is consistent with the categories of tacit knowledge listed in previous section: tacit knowledge not only includes knowledge that is technically hard to articulate, but also knowledge that is social sensitive, informal, etc.

Implications For Helping Secondary Use

What I have reported so far can be summarized as the followings: Inadequate documentation is caused by two reasons. First, data producers are not motivated or not capable of documenting data well for secondary use. Second, documentation is inherently inadequate because of the tacit knowledge problem and communication reduction. When documentation is inadequate, users seek outside information to supplement documentation. However, inadequate documentation is not the only reason that users seek outside information. Users may seek outside information when documentation has no problems. They may seek outside information when their absorptive capacity is not ready for using the data. They obtain outside information about data due to convenient outside information sources and channels. They prefer to obtain information by socializing with people than reading documentation alone. Some users encounter useful information without actively seeking it. When seeking outside information, more than 60% of users turn to various documents, which can be easily included in documentation or at least point to from documentation. More than 60% of users seek information from various people.

These findings above have implications for strategies to help secondary data use. One way to help secondary data use is to improve users' absorptive capacity. Some data archives provide data analysis training for secondary users. Otherwise, we should rely on users' academic training and professional experience to improve their absorptive capacity. A second method to help secondary use is to provide more communication channels between data producers and secondary users, and among secondary users. More communication helps build collaboration relationships, makes outside information immediately accessible for secondary users, increases the chance of encountering information, and facilitates the transfer of tacit knowledge. Currently some data producers provide workshops to train data users and make them familiar with their data. There are also some mailing lists for users of the some datasets. Web 2.0 technologies such as Wiki pages also can be used to facilitate communication. A third method is to provide instructions to data producers and make sure those instructions are known and implemented by data producers. The data archive community has various tools and methods available to help data producers, such as the international standard Data Documentation Initiative (http://www.ddialliance.org/), guidelines provided by various data archives, and the idea that documentation should be incorporated into the whole life cycle of data. But many data producers are not aware of these instructions or do not implement them. Improving the awareness of these tools and methods may help data producers document data better for secondary use. A fourth approach is to provide incentives for data producers who are not motivated to document data well for secondary use. The Coase theorem (Frank, 2007, p. 539 and p. 543) can be applied here. Below are tentative and brief discussions about how Coase theorem can be applied to data sharing.

According to the Coase theorem, when the parties affected by externalities[14] can negotiate costlessly with one another, an efficient outcome results no matter how the law assigns responsibility for damages. When there is a negotiation cost, efficient laws and social institutions are the ones that place the burden of adjustment to externalities on those who can accomplish it with least cost. Here is an example to show what the theory means. A doctor and a candy maker are neighbors. The doctor's ability to examine patients was disturbed by the noise of machinery operated by the candy maker. Suppose the candy maker has access to a soundproofing device that eliminate all noise at a cost of A. the doctor has the option of avoiding the noise by re-arranging his office, which will cost B. If the negotiation of a private agreement between the doctor and the candy maker entails negligible cost, whether the legal system makes the candy maker liable for the noise doesn't affect the efficiency of the outcome for the society. When the negotiation of a private agreement between the doctor and the candy maker is not cost free, if A < B, making the candy maker liable for noise damage is more efficient for the society than otherwise.

If A > B, not making the candy maker liable for noise is more efficient for the society than otherwise. Applying this to data sharing, poor documentation provided by data producers is a kind of noise, which puts data producers and secondary users in the same situation as the candy maker and the doctor. Data producers can solve the problem by taking time and effort to improve the quality of documentation. Secondary users can overcome inadequate documentation by seeking outside information, making compromises, tolerate uncertainties, etc. If the negotiation between the two parties is cost free[15] , an efficient outcome results no matter how the law assigns responsibility for inadequate documentation. In other words, it is not necessary to make data producers liable for inadequate documentation. Making data producers liable does no good for the society versus if we just leave messy data and poor documentation for secondary users to deal with. If the negotiation cost between the two parties cannot be ignored, to achieve higher efficiency for the society, whether we should make data producers liable for poor documentation or not depends on which party can solve the problem with lower cost, which may vary with data producers. Three types of data producers tend not to be motivated to document data well:

1) private companies, for example, Sony Online Entertainment is a company who runs large online gaming environment. It keeps records about the behaviors of game players. But they do not sell those data for profit. Those records can be used by researchers to study some social or psychological issues;
2) government agencies who produce administrative records as by-products of their business process;
3) individuals and small research groups who produce data for their own research.

It may not be cost efficient for private companies and government agencies to spend more effort documenting data for secondary users who tend to be individuals and small research groups. In other words, if only one user or very few users need certain records produced by a government agency or a company. It will be more efficient to not make data producers liable for messy data and poor documentation and let the users take the cost in using data. But if a very large number of users need those data for research, it may be more costly for each individual user to clean up data and take various efforts to supplement documentation. Therefore it may be more efficient to make data producers liable for messy data and poor documentation when the data are wanted by a very large number of users. This actually explains why administrative records are often messy and poorly documented when they are obtained by individual researchers directly from government agencies, but often well documented when they are obtained from an intermediary organization. Administrative data obtained from a government agency by individual researchers are less likely to be used by a large number of users. Administrative records obtained from intermediary organizations are compiled for sharing and intend to be used by a large number of users. It is a different story for data producers who are individuals or small research groups. Users of those data are likely to be individuals or small research groups as well. Which party can overcome the externality of messy data and poor documentation may need to be decided on a case-by-case basis, or depends on how much policy makers value secondary data use. If policy makers value secondary users' cost to overcome poor documentation more than the third type of data producers' cost in improving data and documentation quality, under this value system, it would be more efficient to make data producers liable. Otherwise, it would be more efficient to let secondary users take the cost to use data.

Suppose paying intermediaries to clean up data and improve documentation quality costs less than data producers to do the work themselves, Intermediaries, such as data archives, become a lower cost solution for data producers. Using intermediaries would lead to a more efficient outcome than not using them. Actually, this is exactly what the National Institute of Justice (NIJ) is doing. NIJ requires its grantees to deposit data into a data archive at the end of their grants.

Law enforcement is a coercive force and often involves some sort of punishment for non-compliance. The other way to motivate data producers to improve data and documentation quality is to use rewards. Appropriate rewards can establish mutual benefiting mechanism between data producers and secondary users, and change the situation that even though data producers take effort in preparing data for sharing, the benefit of sharing largely goes to secondary users. Effective punishments forces all data producers without plausible excuses to prepare and deposit data, which would make most data collected under public funds accessible to the public. This gives users' chances to verify the research findings of data producers, which

would deter scientific fraud and misconduct. On the other hand, not all data sets will be used heavily (Niu and Hedstrom, 2007). Under the punishment scenario, even if the data is very unlikely to be used in the future, the data producer still needs to document and share data to avoid punishment. Enforcing uniform strong punishment on all data sets would cause the waste of resources. Unlike the coercive and uniform nature of punishments, rewards are inducive and selective. Rather than forcing researchers, rewards induce researchers to prepare and deposit data. Researchers who expect their data to be used by other people will be motivated to do better in data sharing. Data producers who do not expect their data to be used will not share data, which may be more efficient for the society. In this case, not all federally funded data are made available to the public, chances to deter scientific fraud decreased. Besides, data producers decide their effort in data preparation based on the expected future use of their data, which might be hard to anticipate.

Acknowledgements

This research is funded by the Rackham Research grant for dissertation and the Rackham one-term dissertation writing award from University of Michigan and NSF Award # IIS 0456022. I greatly appreciate the comments and instructions from Professor Margaret Hedstrom.

References

Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

Carlson, S., and Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. Journal of Computer-Mediated Communication, 12(2), article 15. http://jcmc.indiana.edu/vol12/issue2/carlson.html

Collins, H. M. (2001). Tacit Knowledge, Trust, and the Q of Sapphire' Social Studies of Science, 31, 1, 71-85

Fienberg, S. E., Martin, M. E., Straf, M. L. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.

Frank, R. H. (2007). Externalities, Property Rights and the Coase Theorem. Ch. 16 of Microeconomics and Behavior, 6th ed. (McGraw-Hill).

Grudin, J. (1994). Groupware and Social Dynamics: Eight Challenges for Developers.

Communications of the ACM 37/1: 92-105.

Markus, M. L. (2001) Toward a theory of knowledge reuse: type of knowledge reuse situations and factors in reuse success. Journal of Management Information Systems. 18(1), 57-93.

Niu, J. and Hedstrom, M. (2009). Documentation Evaluation Model for social science data: An empirical test. In Proceedings of the Digital Curation Curriculum (DigCCurr) conference, Chapel Hill, NC (April 1-3, 2009).

Niu, J. and Hedstrom, M. (2008). Documentation Evaluation Model for social science data. In Proceedings of the Annual Meeting of the American Society for Information Science Technology (ASIS T), Columbus, OH (October 24 - 29).

Niu, J. and Hedstrom, M. (2007). Streamlining the "Producer/Archive" Interface: Mechanisms to Reduce Delays In Ingest And Release Of Social Science Data. DigCCurr2007 Conference, Chapel Hill, NC (April 18-20).

Polanyi, M. (1962). Personal knowledge: Toward a post-critical philosophy, Harper Torchbooks, New York.

Strathern, M. (2005). Useful knowledge. Lecture presented at The Isaiah Berlin Lecture, Manchester, UK. Not published.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. Strategic Management Journal, 17(Winter Special Issue), 27-43.

Footnotes

[1]Knowledge and information are used interchangeably in this paper.

[2]The Non-parametric Mann-Whitney U-test was used to test the difference.

[5]Professional status means whether the user is a full professor, associate professor, assistant professor, postdoctoral researcher, doctoral student, etc.

[6]Details about the difference between data produced for self-use and data produced for sharing can be found in (Niu and Hedstrom, 2009)

[7]This survey was conducted on grantees of the National Institute of Justice, who are required to

deposit their data to a data archive at the end of their grants.

[9]This N is different from the total number (384) of usable responses to the survey because of missing data. Some users did not answer where they obtained outside information.

[10]This number is smaller than the sum of 66% and 16% because there is overlap between A and B.

[11]We acknowledge that there is some overlap between categories A and E.

[14]The word externality is widely used in economics. It means the impact of one's action on others.

[15]Currently I don't know how the negotiation between different types of data producers and secondary users takes place and how much it costs.