Data sharing in sociology journals

Zenk-Möltgen, Wolfgang; Lepthien, Greta

Data sharing in sociology journals

Wolfgang Zenk-Möltgen and Greta Lepthien

(Data Archive for the Social Sciences,
GESIS – Leibniz Institute for the Social Sciences,
Cologne, Germany)

Abstract

Purpose – Data sharing is key for replication and re-use in empirical research. Scientific journals can play a central role by establishing data policies and providing technologies. This paper analyzes the factors which influence data sharing by investigating journal data policies and the behavior of authors in sociology.

Design/methodology/approach – The websites of 140 sociology journals were consulted to check their data policy. The results are compared with similar studies from political science and economics. A broad selection of articles published in five selected journals over a period of two years are examined to determine whether authors really cite and share their data and the factors which are related to this.

Findings – Although only a few sociology journals have explicit data policies, most journals make reference to a common policy supplied by their association of publishers. Among the journals selected, relatively few articles provide data citations and even fewer make data available – this is true both for journals with and without a data policy. But authors writing for journals with higher impact factors and with data policies are more likely to cite data and to make it really accessible.

Originality/value – No study of journal data policies has been undertaken to date for the domain of sociology. A comparison of authors' behaviors regarding data availability, data citation, and data accessibility for journals with or without a data policy provides useful information about the factors which improve data sharing.

Keywords: Data sharing, Journal policy, Sociology, Empirical social research, Research data infrastructure, Data citation

Classification: Research paper

Introduction

There has recently been considerable growth in attention to issues of data sharing in research (see Kindling et al. 2013; Berman and Cerf 2013). As data is the basis of research in many disciplines, it plays a central role in the scientific method. Central to the scientific method is reproducibility, which can only be achieved by obtaining access to original data. Access can only be provided to research data with good data storage and data preservation. Many funding agencies, data producers, data consumers, or data centers agree that improvements are needed in this area (Feijen, 2011: 29). Replication studies increase the transparency of research and also secure and develop knowledge in a

specific domain. Access to research data is necessary for replication studies. Open and long-term data availability allows knowledge to be verified and makes the research process transparent (see King 1995; Jasny et al. 2011). The secondary use of research data also precludes the performance of redundant studies, provides quality control and can add newly collected data to research datasets. The availability of research data makes it possible to answer new research questions in the same area which may not have been asked by the original authors of a study. In addition, making research data available to others increases the number of citations of the principle investigators which may be regarded as an incentive for the often time-consuming documentation of research data (see Piwowar et al. 2007).

According to the research data life cycle defined by GESIS – Leibniz Institute for the Social Sciences (2014) data sharing is part of phase five, "Archiving and registering" (see Figure 1). This phase also includes the dissemination of archived data to secondary users. The term "cycle" indicates that this last step is also the basis of the first step in the iterative research process. This first step, "Research", is to obtain information about the current state of the art concerning not only existing theories in the literature but also previously collected data on research topics, similar research projects and the institutions involved. As long as no data for secondary analyses are available, the second step, "Study planning", is to design an individual study in accordance with the specified goal and restrictions in order to check the research hypothesis. Data is collected, cleaned and integrated in the next phase, "Data collection", so that it can be used in the "Data analysis" phase. This example from GESIS – Leibniz Institute for the Social Sciences of a research data life cycle is only one of many examples that show the iterative nature of research where the phases of the model are repeated (Wegener et al. 2013). Data availability and data sharing are essential prerequisites to allow for the repetition of the research data life cycle phases.

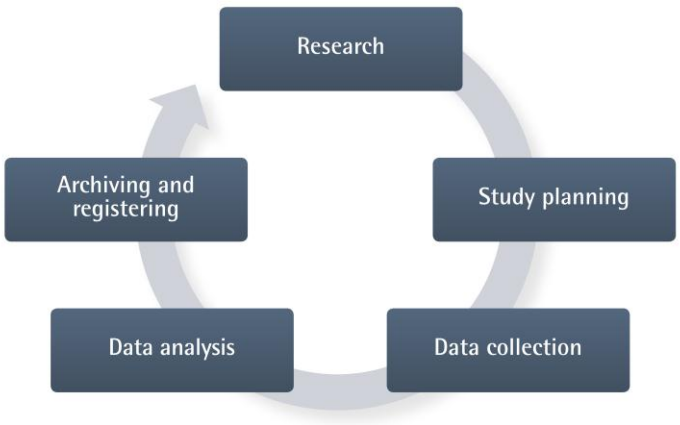

Figure 1: The GESIS research data life cycle (GESIS 2014)

Social science data infrastructures

In the area of social sciences, this service is provided by several longstanding data archives: the Inter-university Consortium for Political Research was founded in 1962 (ICPSR 2014), the GESIS Data Archive for the Social Sciences was founded in 1960 as the Central Archive for Empirical Social Research in Cologne, Germany (Schumann and Mauer 2013). Other examples exist in the Consortium of European Social Science Data Archives (CESSDA) across Europe.[I] Since then, they have been providing services to curate, archive, and disseminate research data in the social sciences. The increasing possibilities which the internet offers for sharing and exchanging data of all kinds are also helping institutions and scientists to build and use data sharing technologies. Examples are the CESSDA Data Catalogue, the Dataverse Network[II], and the recently established data sharing service datorium[III] of GESIS (Linne 2013). However, the availability of effective data sharing technology is just one of the many factors that determine whether scientists really do share their data.

These archives mainly cover disciplines such as sociology, political science, and media research; partly, they also cover some areas of economics, psychology, geography, or health research. Under these circumstances one would expect data sharing in the area of social sciences to be a well established practice. However, studies show that this is not the case: Tenopir et al. found that "most respondents are willing to share at least some of their data (…). Respondents in the sciences are generally more satisfied with current situations and more willing to share than those in disciplines such as medicine or social sciences where human subjects or other restrictions may come into play with some datasets" (Tenopir et al. 2011:14).

Development of data policies

Funders of research, such as the German Science Foundation DFG, the US National Science Foundation NSF or the European Union, have understood that the sharing of research data may increase the scientific value of findings (DCC 2014; Dietrich et al. 2012). In addition, they appreciate that the sharing of research data increases the impact of their investment and avoids unnecessary duplicate expenses. This has resulted in several guidelines which recommend preserving and sharing research data. The NSF "expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work" (NSF 2001:17). The DFG recommends preserving the primary data on which publications are based at the originating institutions for at least ten years (DFG 1998: 12). The German Data Forum (Rat für Sozial- und Wirtschaftsdaten) underlines the importance of services to archive research data and to use metadata standards (RatSWD 2011: 32). The German research council has issued recommendations for data documentation and urges better access to research data (Wissenschaftsrat 2012: 54). The European Union provides a vision in the report "Riding the wave": "All of these principles – our vision – point in the direction of an infrastructure that supports seamless access, use, reuse, and trust of data. It suggests a future in which the data infrastructure becomes invisible, and the data themselves have become infrastructure – a valuable asset, on which science, technology, the economy and society can advance." (EU 2010: 24)

Journals have started to include policies on sharing research data in their submission guidelines for authors. Examples are Science[IV], Nature[V], the American Sociological Review[VI] or Sociological Methodology[VII]. In the domain of political science,

Gherghina and Katsanidou (2013: 9) found that the American Journal of Political Science, European Union Politics, Journal of Conflict Resolution, Journal of Peace Research, Political Analysis, Political Communication, and Politische Vierteljahresschrift (Political Quarterly) all have policies which are explicit in expecting authors to share data. With regard to the discipline of economics, Vlaeminck and Siegert (2012: 12) found that the data availability policy of the American Economic Review represents best practice that is re-used by other journals in the same field. This paper tries to answer similar questions for the domain of sociology, where no study has been conducted so far.

Studies investigating journal data policies

Gherghina and Katsanidou (2013) have examined data availability policies in political science by analyzing a representative sample of 120 ranked political science journals (following the 2010 Social Sciences Citation Index of Thomson Reuters). They found that 18 (15%) of the 120 journals have adopted data availability policies and seven (5.8%) plan to adopt them. Their analysis reveals that an increasing number of political science journals are adopting data policies. They also tested the correlations between the existence of data policies and journal characteristics, such as age of the journal, frequency of issues, language of the journal, type of audience, and impact factor. The strongest correlation was observable for the impact factor: journals that are cited more often are more likely to have a data availability policy than publications with fewer citations (Gherghina and Katsanidou 2013).

Vlaeminck and Siegert (2012) have investigated the role of research data for economics journals. They surveyed 141 journals and ascertained that 20% (29) of the journals have data policies which require data and other materials to be sent to the editors, who can then make them available to third parties. They also found that journals which have a data availability policy have a higher impact factor than journals without such a policy. This means that primarily high ranked journals implement these guidelines. One reason for this is that these journals have the infrastructural resources that are necessary for the technical implementation of such a policy.

In a second step Vlaeminck and Siegert analyzed the content of the data availability policies based on 9 content-related criteria, such as "the guidelines should be mandatory" or "authors should transmit the data sets they used to the editors". They found that 82.8% of the data availability policies were mandatory for the authors and almost 90% of the examined policies included submission of data sets to the editors. Furthermore, 51.7% of policies required the submission of a calculation code (data cleaning or analysis syntax), 65% required descriptions of the data provided and 62% the submission of programs written by the authors.

Documenting and providing data is time-consuming and to date has not been adequately remunerated by the scientific system. Thus it can be assumed that the prospect of publication in a prestigious, high ranked journal will offer a greater incentive for researchers to document and share their data (Huschka et al. 2011).

It is not immediately apparent whether this also applies in the field of sociology, and this is why we have conducted this small study of sociology journals and the behavior of their authors.[VIII]

Research questions and research design

The first area of interest covers the behavior of journals' editorial boards with regard to data policy: Our aim was to investigate whether sociology journals already have a data policy for authors in place and, if so, how detailed and mandatory the rules of such policies are. We also investigated whether the existence of a data policy for a journal correlates with its impact factor, language, number of issues, or the age of the journal. The findings also reveal the extent to which differences exist between the field of sociology and other domains, such as political science or economics.

The second area of interest is the behavior of authors. We were interested in determining whether the authors of articles in sociology journals state that the data used to produce an article's results are available to others. Our hypothesis is that a difference in this respect will be found between authors contributing to journals which have a data policy and those writing for journals which do not have such a policy. We also wished to determine whether authors really do cite the data that they have used and, where this is the case, whether they do so in the way recommended by the journal. To confirm, we tried to determine whether the cited data was really available to us. The cited data may not really be accessible for a number of reasons, e.g. outdated URLs or insufficient information to identify or find the dataset. This is important to investigate because in these cases replication would not be possible.

In the third area we wished to determine whether there are any correlations between data availability statements, citations, and accessibility in the articles on the one hand, and journal attributes regarding data policy, impact factor, language, number of issues, or age on the other, the aim being to determine whether general expectations concerning data sharing can be fulfilled in the domain of sociology.

Methodology

First, we chose a comprehensive list of research journals from the Social Science Citation Index (SSCI) in the Web of Science (Thomson Reuters 2013). 140 journals are listed in the area of sociology (general, no specific topics), 135 of them with an impact factor. We visited the websites of all the journals to review the author guidelines and search for a data policy. The availability of a data policy was coded in the categories "no", "should", and "yes". Some journals have a dedicated policy developed internally and which was coded "yes"; others do not have a policy at all and this was coded "no". The third category "should" reflects the fact that some publishers or associations have general recommendations for all their journals. These recommendations, which are somewhat weaker than dedicated journal policies, need to be taken into account in the analysis. Following Gherghina and Katsanidou (2013), we also coded the age of the journal, issues per year, language (English, other or both) and the ISI Web of Science impact factor. We calculated correlation coefficients between all these factors to identify related attributes.

Second, we selected relevant journals in the fields of sociology in order to review the articles published in these journals. According to the ISI Web of Science, we selected both high and low impact factor journals. Furthermore, we selected both journals which have a data policy and journals which do not. The selection also included some German speaking journals in order to determine whether there is a difference between articles

published in international and in German journals. We coded each article according to whether it is an empirical article based on data, a theoretical article that does not use data, or of an entirely different nature, e.g. a book review, comment, editorial note or announcement. We then coded three variables for each empirical article (with "yes" or "no"): First, does the article state that the data is available? Second, could we prove that the data is really accessible? Third, does that article say that the data was collected by the authors themselves and that the data was available from them? The reliability of the coding was confirmed by an independent coding of a small sample of the selected articles by a different person and no deviations were found.

From the list of 140 journals we selected the American Sociological Review (ASR) as an example of a journal with high impact factor (ISI 4,077 – second rank) and English language. The American Journal of Sociology (AJS) (ISI 3,414 – fourth rank) and the Sociological Methodology (SM) (ISI 3,167 – sixth rank) also fall into this category. As examples of lower impact factor and German language journals we selected the Zeitschrift für Soziologie (ZfS) (0,604 – rank 88) and the Kölner Zeitschrift für Soziologie und Sozialpsychologie (KZfSS) (0,481 – rank 96). We analyzed all the 581 articles published in these journals between 2012 and 2013.

In addition we combined the information about the journals with the information about the articles. The attributes of the journals that were collected in the first step were added to the data about journal articles collected in the second step. Correlation coefficients were calculated to see whether any relationships exist between journal attributes of data policy, impact factor, language, number of issues, and age of the journal with author's statements and behavior concerning data availability statements, data citation, and data accessibility.

Findings

Of the 140 journals, 122 (87.1%) are English language journals, 14 (10%) use other languages, 4 (2.9%) are bi-lingual. A majority of journals – 86 (61.4%) – appear four times a year, 19 (13.6%) have fewer issues per year and 35 (25%) have more issues (up to 12, but mostly six per year). The age of the journals ranges from six years to 118 years, with a mean value of 38 years (one missing value). The first quartile of the journal age is at a value of 22 years, the second quartile is at 36 years, and the third quartile is at 51 years.

Data policies of sociology journals

Seven journals (5%) were found to have an explicit data policy: American Sociological Review (ASR), Sociological Methodology (SM), Sociological Theory (ST), Sociology of Education (SE), Social Science Quarterly (SSQ), Contemporary Sociology (CS), and Teaching Sociology (TS). Apart from Social Science Quarterly, all these journals are published by the American Sociological Association (ASA). The ethical standards of this association have the following to say about data sharing:

"Sociologists make their data available after completion of the project or its major publications, except where proprietary agreements with employers, contractors, or clients preclude such accessibility or when it is impossible to share data and protect the confidentiality of the data or the anonymity of research participants (e.g., raw field notes

or detailed information from ethnographic interviews)."(American Sociological Association 2014)

Another 94 journals (67.1%) refer to a common policy provided by an association of their publishers, the Association of Learned and Professional Society Publishers (ALPSP). On the website of this association it is stated that "data sets, the raw data outputs of research, and sets or sub-sets of that data which are submitted with a paper to a journal, should wherever possible be made freely accessible to other scholars" (Association of Learned and Professional Society Publishers 2006). This was coded as "should" for the existence of a data policy. Together with the seven journals with an explicit data policy, 101 sociology journals (72.1%) recommend that their authors deposit and share the research data used in their articles. 39 (27.9%) of the sociology journals do not have a data policy in place.

Journal characteristic	Data policy: yes	Data policy: yes or should	No. of cases
Impact factor	0.204*	0.384**	135
English language	0.088	0.618**	140
Issues per year	0.005	0.268**	140
Age of journal	-0.045	-0.031	139

Table 1: Correlations between existence of data policies and journal type (Spearman rho, * indicates significance level of 0.05, ** indicates significance level of 0.01)

We have used four characteristics to analyze the correlations between journal characteristics and data policies: impact factor, English language, issues per year, and age of the journal. For the impact factor of SSCI we used the five year average and, if not available, the 2013 value. English language was coded as one, other language or bi-lingual was coded as zero. Issues per year and age of the journal were calculated for 2013. We used Spearman's rho as the correlation coefficient because of the dichotomous or ordinal character of the variables and because no linear relationship can be assumed (however, the Pearson correlations show the same trends).

The results (see Table 1) show that a higher impact factor is positively correlated with the availability of a dedicated data policy, and that this correlation is even stronger and more significant for the availability of a dedicated or common data policy. The use of English language is not correlated to the presence of a dedicated data policy, but does show the highest correlation of the analyzed variables for the dedicated or common data policy. This is due to the fact that a very high number of the journals follow the recommendations of the ALPSP and are English language publications. Also the number of issues is not correlated to the presence of a dedicated data policy, but to the presence of a dedicated or common data policy. Because the majority of journals with a dedicated or common policy belong to ALPSP, the significant correlations with this factor may indicate attributes of ALPSP member journals. The age of the journal is not correlated to either of the data policies.

Data sharing in sociology articles

For the second area of interest we focused on articles in selected journals. Articles were selected for analysis by choosing five journals with different characteristics (see Table 2). Among the journals, AJS and ASR are two high impact factor journals, SM is medium, ZfS and KZfSS have comparably lower impact factors. ZfS and KZfSS are German language journals, the others use English. SM only has one issue per year, KZfSS has four issues per year, and the others have six issues per year. AJS is the oldest journal in the dataset, KZfSS and ASR also have a long history, SM and ZfS are medium aged. Only ASR and SM actually have a dedicated data policy in place.

Journal	Impact factor (5-year)	Language English	Issues per year	Age in years	Data Policy
ASR	5.563	Yes	6	77	Yes
AJS	5.239	Yes	6	118	No
SM	2.662	Yes	1	42	Yes
ZfS	0.724	No	6	41	No
KZfSS	0.602	No	4	92	No

Table 2: Characteristics of selected journals in sociology

We analyzed 581 articles published in issues of the selected five journals in 2012 and 2013. 328 or more than half of the articles (56.5%) were published in AJS, 89 (15.3%) in ASR, 67 (11.5%) in KZfSS, 59 (10.2%) in ZfS, and 38 (6.5%) in SM. 41 of the articles (7.1%) did not use any empirical data, another 318 articles (54.7%) were book reviews, editorial notes, comments, or announcements (figures are not shown in the table). The remaining 222 articles (38.2%) based on data were consequently considered to be empirical. The rate of empirical articles differs a great deal between the selected journals: ASR has only a few articles that are not empirical, AJS and SM have only a small rate of articles based on empirical data (see Table 3).

The empirical articles in these journals are the ones that should make data available to others for replication and re-use, and we therefore only used these articles for the further examinations in this paper.

We analyzed whether the authors of the empirical articles state, either in the introduction, in a footnote or in the references, whether their data is available. In a second step we tried to find the data itself in order to determine how accessible it actually is. This involved checking the URL or persistent identifier like DOI that were reported in the article. However, we did not actually order or download the data. As some authors do not deposit their data at a publicly accessible location, the third step was to look explicitly at whether the authors stated that the data had been collected by themselves and could be obtained via their contact address.

Journal	Articles	Empirical	Data available	Availability proven	Own data & available
AJS	328	55 (16.8%)	19 (34.5%)	5 (9.1%)	1 (1.8%)
ASR	89	81 (91.0%)	61 (75.3%)	25 (30.9%)	5 (6.2%)
SM	38	9 (23.7%)	1 (11.1%)	1 (11.1%)	-
ZfS	59	38 (64.4%)	12 (31.6%)	1 (2.6%)	1 (2.6%)
KZfSS	67	39 (58.2%)	22 (56.4%)	3 (7.7%)	-
Sum	581	222 (38.2%)	115 (51.8%)	35 (15.8%)	7 (3.2%)

Table 3: Characteristics of articles published in selected journals (percent base for rate of empirical articles are all articles, for other columns the percent base is the number of empirical articles only)

Figures show (see Table 3) that just over half of the articles state that the data are available (51.8%). ASR has the highest rate of articles that state that data is available (75.3%); ASR also has a large number of empirical articles as well as a dedicated data policy in place. The other journal with a dedicated data policy, SM, has only a very low rate (11.1%) of articles for which data availability is stated. Even bearing in mind that SM publishes very few empirical articles, this is an astonishingly low value. The journals with no data policy in place also have a considerable number of articles that state that data is available: KZfSS has 56.4%, AJS 34.5%, ZfS 31.6%. This shows that the availability of data in published articles does not depend critically on whether the journal has a data policy in place or not. Most authors seem to agree that it is important that the data underlying their published articles should be available: over 50% state in their articles that the data is available. On the other hand, even journals which actually have a data policy in place do not automatically ensure that a statement about data availability is included in the articles. This is surprising considering how much importance researchers generally attach to proper data citation (Tenopir et al. 2011: 11) when their data is used by others.

Replication of published analyses depends very much on the ability to access the data that has been used. For this reason we tried to find the data used in articles in those cases where the authors stated that data was available. We followed the URL or persistent identifiers like DOI that were mentioned in the article to see whether we could identify the actual dataset and find out about access restrictions. Even though we were aware of common complaints about frequently changing URLs, we were surprised how few of the purportedly available datasets used in the articles we were able to find: The highest rate for articles with data proven to be available was 30.9% in ASR. For 81 empirical articles in ASR, 61 (75.3%) had mentioned that the data was available, but only 25 (30.9%) could be proven to be accessible. The rate is even lower for other journals: For SM, the other journal with a data policy in place, out of nine empirical articles only one (11.1%) mentioned that the data was available and this could also be confirmed. AJS has five of 55 (9.1%) empirical articles with proven data availability, KZfSS has three of 39 (7.7%), and ZfS has only one of 38 (2.6%). Overall, of 222 empirical articles only 115 (51.8%) stated that the data is available, and only 35 (15.8%) could be confirmed to have the associated research data accessible.

Because only so few datasets are available for replication analyses, we wanted to know whether authors work with data of their own that may not be accessible to the public and for which therefore no URL or persistent identifier may be available. If this was the case one would expect to be notified of this in the article itself and about the possibility of obtaining the data directly by contacting the author. After performing such an analysis we found five articles (6.2%) in ASR and one each in AJS (1.8%) and ZfS (2.6%) where this was the case. In a broader sense one might consider these as being articles for which data is available. The data policy of ASR does appear to encourage statements about data availability, but the actual rate of datasets that are really available for replication purposes for the analyzed journals remains low. This means that most of the empirical datasets used in published articles in the analyzed journals are still not accessible for replication or re-use analyses.

Author behavior and journal characteristics

The third area of interest we investigated was the connection between journal characteristics and author behavior. To this end we calculated the correlations for the empirical articles in the selected journals between the three categories of stated data availability, proven data availability, and use of own data, on the one hand, and journal characteristics of impact factor, language, age, number of issues, and available data policy on the other.

A positive correlation (see Table 4) can be found between data availability as stated in an empirical article and the existence of a data policy for the journal. This also correlates with a high impact factor of the journal. The same correlations can be found for the proven data availability of articles, with even larger correlation coefficients. In addition, the use of English language is also positively correlated with proven data availability. The existence of a data policy for the journal, a higher impact factor, and the use of English language increases the chance that the data used by the article will be accessible for replication analysis and re-use. For the case of authors stating that they use their own data and make them available no significant correlations were found.

Journal characteristic	Data available	Availability proven	Own data & available
Impact factor	0.244**	0.284**	0.126
English language	0.112	0.211**	0.077
Issues per year	0.056	0.105	0.094
Age of journal	-0.032	-0.044	-0.053
Data policy	0.282**	0.297**	0.114

Table 4: Correlations between articles mentioning data availability and journal characteristic (Spearman rho, ** indicates significance level of 0.01, N=222 empirical articles)

Discussion

Data policies of sociology journals

For our first research question we were able to show that nearly three quarters of the ranked journals in the domain of sociology have a data policy for authors in place that requires data availability. Very few sociology journals have a policy which applies only to the journal itself; most journals refer to the common policy of their publisher's association (ALPSP). In the light of the recommendations to increase data sharing this can be seen as a very good basis for future replication projects. However, most of the data policies are very general and do not provide much detail about the contents and technologies that should be used to share research data. They also make many exceptions for cases where the data might not be shared by authors. This leads to a situation in which data sharing may be seen by authors as a recommendation only rather than as mandatory.

The existence of a data policy for journals is positively correlated with higher impact factors of the journals. If we take into account the data policies of journals from their publisher, it is also positively correlated with higher impact factors, the use of the English language, and the number of issues per year. More professional journals with more readers and more contributions appear to be forging ahead in the setting of standards for data sharing in the field of sociology. One reason for this may be that these journals can exert more influence on their authors. Publication in a high impact journal rewards the greater effort involved in providing research data. These journals are also seen as high quality publications, whereby the availability of research data for replication is one indication of their quality.

A comparison of our findings in the area of sociology with studies of other domains, such as political science (Gherghina and Katsanidou 2013) or economics (Vlaeminck and Siegert 2012), reveals some interesting differences. The rate of journals in the field of political science that do have a data policy is somewhat higher (15%) than in the field of sociology (5%). The same is true of the rate of economic journals with data policy with 20%. This picture is very different, however, when compared with the 72.1% of sociology journals that either have their own policy or a common policy: The rate of journals with some form of data policy in sociology is much higher than in political science or in economics. The reason for this is obviously that a lot of journals in sociology follow the recommendation of a data policy by their association of publishers ALPSP.

We also found that journals with high impact factor are more likely to have a data availability policy – here we have the same result for sociology as has been found for political science and economics. The use of English language has been found to be correlated with data availability policies in sociology and political science, the correlation of other characteristics, such as issues per year and age of journal, differs in the findings (these characteristics have not been investigated for economics). Altogether, these findings support the view that it is high impact factor journals which install a data availability policy for their authors and help in supporting replication analysis.

Data sharing in sociology articles

The second area of interest was the behavior of the authors of empirical articles in the selected sociology journals. We were able to show that about half of them state in their

articles that the data is available or cite the data they use, and that the rate differs a great deal between different journals. Considering that we selected quite different characteristics of journals it was interesting to see that not only high impact journals or English language journals have a high rate of articles with data availability stated.

When analyzing the real availability of the research data, only about 16% of the dataset from all empirical articles could be considered accessible. The highest rate of accessible research datasets of the analyzed journals was about 30% – less than one third of all empirical articles. This increased only marginally when we added the numbers of articles in which authors state that they used their own data and are willing to make this data available. In the field of sociology this shows that expectations for the comprehensive availability of research data for replication analyses are far from being fulfilled.

Author behavior and journal characteristics

In the third area of analysis we could show that correlations exist between data availability statements or data citations and the characteristics of journals: First, the existence of a data policy was found to be positively correlated with data availability and data accessibility. This is the strongest correlation, but it is not strong enough to determine whether authors state or guarantee data availability. Second, a high impact factor was almost equally strongly correlated with data availability and data accessibility. Third, the use of the English language in the analyzed journals was also found to be correlated with the accessibility of data used in the articles. These relationships reveal a picture of a data sharing culture in sociology that has already developed, even if it is not as yet a mainstream culture. Despite these correlations, the percentage of articles with proven data availability is not only low in general, but also low in journals with a data policy. The quite low number of available datasets from empirical articles is especially disappointing given that data sharing infrastructures and technologies in social science have now been in existence for quite some time.

Limitations and open issues

Our analysis is limited by the fact that it drew only on selected journals in the general category of sociology in the SSCI. There are many sociology journals listed in specific topic categories (more than 3000 journals altogether) which were not analyzed. It is not clear whether the journals selected are representative of the complete area of sociology. Another limitation is that only articles from five journals have been analyzed for the behavior of authors. Even if the selection consists of a considerable amount of articles from a diverse variety of sociology journals, systematic bias in the selection of these articles cannot be excluded completely. However, the higher ranked journals tend to be trend setters for the rest of the journals. Thus we expect to find similar results if we run a wider-ranging analysis.

Further investigation could be done by analyzing the statistical evidence presented in the articles and the reporting of errors to see whether this has a bearing on willingness to share data. This has already been undertaken for the domain of psychology by Wicherts et al. (2011) who found "the reluctance to share data to be associated with weaker evidence (…) and a higher prevalence of apparent errors in the reporting of statistical results" (Wicherts et al. 2011: 1). There might also be a connection with the low quality data management and data documentation.

Conclusions

The overall findings of our study show that good progress is being made in specific areas of sociological research with the sharing of data from articles published in sociology journals. However, some important gaps have been shown with some low rates of data availability statements or data citations and missing data accessibility in some of the journals. Data sharing might be encouraged by adopting highly professional standards, such as establishing explicit data policies, using policies that other journals or publishers already have developed, and providing easy and understandable means of data sharing. This could also be achieved by recommending data repositories and data archives, which provide services to identify datasets by persistent identifiers, curating the datasets, and providing advice and access to research datasets. We have shown that available data sharing technologies alone are not sufficient to foster data sharing in sociology. In addition, rules for data sharing have to be adopted and made more effective. If the data policies of journals were checked on a regular basis or if they were made mandatory in a more formal way, data sharing could become second nature for the authors of empirical articles. Policy recommendations of funders and science foundations target high quality science where replication and re-use is one major claim – but this quality of science still needs to be delivered.

References

American Sociological Association (2014), "Ethical Standards. Sections 12 - 20.06", available at http://www2.asanet.org/members/ecostand2.html#12 (accessed 5 May 2014).

Association of Learned and Professional Society Publishers (2006), "Databases, data sets, and data accessibility – views and practices of scholarly publishers. A statement by the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical and Medical Publishers (STM).", available at http://www.alpsp.org/Ebusiness/AboutALPSP/ALPSPStatements/Statementdetails.aspx?ID=55 (accessed 8 May 2014).

Berman, F. and Cerf V. (2013), "Who Will Pay for Public Access to Research Data?", *Science*, Vol. 341, No. 6146, pp. 616-617.

Deutsche Forschungsgemeinschaft (1998), "Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission ‚Selbstkontrolle in der Wissenschaft'", available at: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (accessed 20 February 2013).

Dietrich, D., Adamus, T., Miner, A. and Steinhart, G. (2012), "De-Mystifying the Data Management Requirements of Research Funders", *Issues in Science and Technology Librarianship*, No. 70, DOI:10.5062/F44M92G2, available at http://www.istl.org/12-summer/refereed1.html (accessed 30 June 2014).

Digital Curation Centre (2014), "Overview of funders' data policies", available at: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies (accessed 30 June 2014).

European Union (2010), "Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. A submission to the

European Commission", available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (accessed 20 February 2013).

Feijen, M. (2011), "What researchers want. A literature study of researchers' requirements with respect to storage and access to research data", available at: http://www.surf.nl/nl/publicaties/documents/what_researchers_want.pdf (accessed 7 February 2013).

GESIS (2014), "Services for the Social Sciences", available at: http://www.gesis.org/en/services/ (accessed 7 April 2014).

Gherghina, S. and Katsanidou, A. (2013), "Data Availability in Political Science Journals", *European Political Science*, Vol. 12, No. 3, pp. 333-349.

Huschka, D., Oellers, C., Ott, N. and Wagner, G.G. (2011), "Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften", working paper 184, Working Paper Series des Rates für Sozial-und Wirtschaftsdaten, Berlin.

ICPSR (2014), "Timeline", available at http://www.icpsr.umich.edu/icpsrweb/content/fifty/history/timeline.html (accessed 8 April 2014).

Jasny, B. R., Chin, G., Chong, L. and Vignieri, S. (2011), "Again, and Again, and Again …", *Science*, Vol. 334, No. 6060, p. 1225.

Kindling, M., Simukovic, E. and Schirmbacher, P. (2013): "Forschungsdatenmanagement an Hochschulen – Das Beispiel der Humboldt-Universität zu Berlin", available at: http://edoc.hu-berlin.de/libreas/23/kindling-maxi-1/PDF/kindling.pdf (accessed 5 May 2014).

King, G. (1995), "Replication, Replication", *Political Science and Politics*, Vol. 28, No. 3, pp. 443-499.

Linne, M. (2013), "Sustainable data preservation using datorium: facilitating the scientific ideal of data sharing in the social sciences.", in Borbinha, J., Nelson, M., Knight, S. (Ed.): *Proceedings of the 10th International Conference on Preservation of Digital Objects, Lisbon*: Biblioteca Nacional de Portugal, pp. 150-155.

National Science Foundation (2001): National Science Foundation Grant General Conditions. http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf (accessed 7 April 2014).

Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007), "Sharing Detailed Research Data Is Associated with Increased Citation Rate", *PLoS ONE*, Vol. 2, No. 3, pp. e308.

RatSWD (Ed.) (2011), "Auf Erfolgen aufbauend. Zur Weiterentwicklung der Forschungsinfrastruktur für die Sozial-, Verhaltens- und Wirtschaftswissenschaften. Empfehlungen des Rats für Sozial- und Wirtschaftsdaten", Budrich, Opladen & Farmington Hills.

Schumann, N. and Mauer, R. (2013), "The GESIS Data Archive for the Social Sciences: A Widely Recognised Data Archive on its Way", *International Journal of Digital Curation*, Vol. 8, No. 2, pp. 215-222.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., et al. (2011): "Data Sharing by Scientists: Practices and Perceptions", *PLoS ONE*, Vol. 6, No. 6, pp. e21101, doi:10.1371/journal.pone.0021101

Thomson Reuters (2013), "Social Sciences Citation Index. Sociology. Journal List", available at http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jlresults.cgi?PC=SS (accessed 19 November 2013).

Vlaemick, S. and Siegert, O. (2012), "Welche Rolle spielen Forschungsdaten eigentlich für Fachzeitschriften? Eine Analyse mit Fokus auf die Wirtschaftswissenschaften", working paper 210, Working Paper Series des Rates für Sozial-und Wirtschaftsdaten, Berlin.

Wegener, D., Baran, E., Zenk-Möltgen, W. and Zapilko, B. (2013), "Towards integrating the research data life cycle of the social sciences based on semantic technology", in Horbach, Matthias (Ed.), *Lecture Notes in Informatics, Proceedings of Informatik 2013 - Informatik angepasst an Mensch, Organisation und Umwelt, 16-20 September 2013, Koblenz, Germany*, GI-Edition P-220, Gesellschaft für Informatik, Bonn.

Wicherts, J.M., Bakker, M. and Molenaar, D. (2011), "Willingness to Share Reseach Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results", *PLoS ONE*, Vol. 6, No. 11, pp. e26828, doi:10.1371/journal.pone.0026828

Wissenschaftsrat (2012), "Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020", available at http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf (accessed 20 February 2013).

[I] http://www.cessda.net/
[II] http://thedata.org/
[III] https://datorium.gesis.org/
[IV] http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail
[V] http://www.nature.com/authors/policies/availability.html
[VI] http://www.sagepub.com/journals/Journal201969/manuscriptSubmission
[VII] http://www.sagepub.com/journals/Journal201996#tabview=manuscriptSubmission
[VIII] The data of this study are available from datorium: Zenk-Möltgen, W. and Lepthien, G. (2014), "Data from: Data sharing in sociology journals", Version 1.0, Dataset, GESIS datorium, doi:10.7802/65