Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

DASISH Work Package 6: "Legal and Ethical Issues"

Deliverable: D6.2
Deliverable Name: "Sample merged paradata sets"
Deadline: 30th June 2013
Nature: Demonstrator

Responsible: MEA (MPG)
Work Package Leader: MEA (MPG)

Contributing Partners and Editors: Daniel Schmidutz, MEA (MPG)
Johanna Bristle, MEA (MPG)

Sample Merged Paradata Sets:

Ethical and Legal Issues

CONTENT

1 Introduction

The deliverable "Sample merged paradata sets" of work package 6 "Legal and Ethical Issues" (WP6) of the "Data Service Infrastructure for the Social Sciences and Humanities" (DASISH) project deals with an important preparatory step towards the analyses of confidential paradata that is generated in the process of survey production.

In the process of producing survey data much paradata, i.e. data about the process of survey production, are generated. Especially with the increasing use and further development of technological means in the context of survey-based data collection, such as computer-assisted personal interviewing (CAPI) techniques and with the implementation of web surveys, the amount of information on the process of survey production has increased:

> "Respondents in web surveys leave electronic traces as they answer survey questions, captured through their keystrokes and mouse clicks. In telephone surveys, automated call scheduling systems record the date and time of every call. In [computer-assisted] face-to-face surveys, interviewers' keystrokes are easily captured alongside the interview and so are audio or even video recordings of the respondent-interviewer interactions. Each of these is an example of paradata available through the computerized survey software." (Kreuter, 2013: 2)

1.1 Scientific and Methodological Value of Paradata

However, not only with regard to the collection of paradata a rapid growth can be observed in the recent years – nowadays, survey researchers are also increasingly making use of paradata, such as keystroke data or contact protocols. According to Couper and Singer paradata are used "to evaluate and improve survey instruments but also to understand respondents and how they answer surveys" (Couper and Singer, 2013: 57).

Paradata are key data for analysing data quality in survey production. Most commonly they are used during survey production for monitoring the fieldwork, including the evaluation of interviewer performance, and the data production process. If up-to-date paradata is available it can be used for implementing responsive designs to guide data production efficiently and improve data quality (Groves and Heeringa, 2006). In general, paradata is used for understanding and improving survey management.

One key indicator often used for determining the quality of survey data is response rates. Since there is a trend that response rates are decreasing worldwide, and especially in Europe, it is important to put more effort into understanding nonresponse and response patterns. For these analyses, survey methodologists mainly rely on paradata. In this connection, there also is a strong demand from the survey methodology community to make paradata of surveys available (e.g. Kreuter, 2013).

Besides this major area of using paradata which aims to improve data quality as well as the entire process of survey production and also intends to make the survey production more transparent to data users and the survey data community, paradata also is used "to describe and classify response behavior […] or to relate response behavior to data quality." (Heerwegh 2002: 2).

1.2 Legal and Ethical Aspects

Currently, "[one] obstacle to releasing paradata are unclear legal and ethical considerations." (Kreuter 2013: 8). Existing codes of ethics are not very clear on the issue of paradata (Couper and Singer 2013), and also from a legal perspective it is in many cases not clear under which conditions paradata should be collected and how they may be used and released.

According to Couper and Singer, "[s]ince the introduction of paradata, researchers have been asking whether and how respondents should be informed about the capture and use of their paradata while completing a survey" (Couper and Singer, 2013: 57). Furthermore, some of the paradata collected may be of a sensitive nature and therefore need close ethical and legal consideration. The extent to which different types of paradata impose new legal and ethical challenges to the survey researchers, including new and special data protection requirements, requires attention, not least due to the legal and ethics issues and associated procedures to which they give rise in the day-to-day operation of survey production.

2 Scope and Objectives

2.1 Overall Objectives of Work Package 6

WP6 addresses various legal and ethical issues that modern research in the SSH is confronted with. Following the "Description of Work" (DoW), Annex 1 to the Grant Agreement of the DASISH project, WP6 has the following main objectives:

- To identify the legal and ethical issues, constraints and requirements for all data types occurring in the SSH domain as result of data integration and linking,
- To cope with legal and ethical challenges imposed by the new data types emerging in the social sciences and humanities,
- To look for professional long-run preservation strategies and policy-rules that can be applied to data collections in the social sciences and humanities.

2.2 Paradata-related "New Ethical and Legal Challenges"

Task 6.1 of WP6 of the DASISH project particularly is concerned with data types imposing "New Ethical and Legal Challenges" to the social sciences and humanities (SSH). It concentrates on the identification of new ethical challenges and legal requirements related to the various data types being recorded in modern research in the social sciences and humanities (SSH).

This demonstrator focuses on the compilation and linkage of paradata, using different data sources from SHARE. However, as part of Task 6.1, it also aims to identify legal and ethical issues connected to the collection and use of paradata and addresses challenges that need to be solved when collecting, processing and offering these data to a wider user group in a cross-country scenario.

While paradata themselves cannot be considered as a new type of data in the field of population-based survey research, obviously "a more structured approach in choosing, measuring, and analyzing key process variables is indeed a recent development" (Kreuter, 2013: 2; cf. Couper and Lyberg, 2005). In a sense, paradata only recently attracted the full attention of researchers conducting field surveys when realising the scientific and methodological value of this data. However, since on the one hand "[t]he number of surveys that collect and provide paradata is growing quickly, and [...] new applications and monitoring systems are develop[ed currently]" (Kreuter, 2013: 8), while on the other hand legal and ethical issues remain unclear, it becomes increasingly important to systematically investigate the ethical aspects and the legal requirements related to different types of paradata.

In WP6, besides the "Report about new IPR Challenges", which also addresses legal and ethical issues related to the collection and the use of paradata in the context of transnational survey research, special attention is given to this topic in the context of this deliverable and deliverable D6.3 ("Exemplary analyses of confidential paradata", due month 36). Here WP6 closely cooperates with WP3 ("Data Quality") with regard to the compilation of a merged paradata set making use of existing data sources from SHARE and subsequently with regard to the analyses of paradata that require special legal and ethical considerations.

The outcome of D6.2 will feed into D6.3 ("Exemplary analyses of confidential paradata") and also in Task 6.2 (i.e. the "Virtual L&E Competence Centre" and the "Handbook on legal and ethical issues for SSH data in Europe").

3 Definition and Differentiation of Paradata

Up to now, there is no standard definition of paradata. Originally, the term paradata referred to computer-generated data about the process of survey data collection (Couper 1998). Paradata was understood as a by-product of survey production such as keystroke data, time stamp data or audit trials. More recently, a broader concept of paradata has become common, which also includes call record information, interviewer observations and information on the interviewers (Couper and Lyberg, 2005; Kreuter and Casas-Cordero, 2010). The "term paradata has been expanded to include a broad range of auxiliary data on the survey process" (Couper and Singer, 2013: 57).

For the purpose of this deliverable we refer to the broader paradata concept and define paradata as micro-level data about the process of survey production. At this, our working definition of paradata is limited to micro-level data in order to be able to distinguish between paradata and metadata, which usually are described as 'data about data'. While, according to Kreuter, metadata "can be seen as macro-level information about survey data" (Kreuter, 2013: 3), such as information about the sampling frame, sampling methods, variable or value labels, percentage of missing data per variable, "[p]aradata capture information about the data collection process on a more micro-level [even though s]ome of this information forms metadata if aggregated" (Kreuter, 2013: 3). For example, keystrokes that capture the minutes needed to interview a respondent or even the time needed for a specific module or a single question, forms information on the average length of the interview or a specific module etc. if aggregated. Although there is a certain overlap between these two definitions of paradata and metadata, there is a crucial difference with regard to the granularity.

Furthermore, our definition of paradata also covers certain types of so-called 'auxiliary data'. Even though researchers frequently refer to this term, in accordance with Kreuter, it is noted that "the definition of this term has not quite been settled upon. The keyword auxiliary data has been used to encompass all data outside of the actual survey data itself, which would make all paradata also auxiliary data. Also contained under auxiliary data are variables from the sampling frame and data that can be linked from other sources" (Kreuter 2013: 3-4). In order to avoid such overlaps and confusions in the context of this deliverable, the term auxiliary data will only be used in relation to additional data obtained separately from external sources (i.e. data not collected in the course of the original survey) or with a specifically targeted effort (i.e. data captured through additional systems).

Accordingly, a differentiation between two types of paradata will be made[1]:

[1] The definition of auxiliary data as well as the differentiation between 'process paradata' and 'auxiliary paradata' has been modelled after Kennickell, Mulrow and Scheuren, 2009.

a) Process paradata: data about the process of survey production recorded as a by-product in the course of conducting a survey, such as listing information (day, time, edits), keystrokes (response times, interview length, back-ups, edits, edit-failures), contact data (day, time, outcome) and gross sample data.

b) Auxiliary paradata: additional data about the process of survey production obtained separately from external sources or with a specifically targeted effort to enhance the information on the survey production process, such as interviewer observations (sample unit characteristics), information on the interviewers (interviewer demographic characteristics), external supplementary data about the sample cases (e.g. geo-referenced data).

4 Sample Merged Paradata Sets in SHARE

In SHARE four different types of paradata are collected and processed:

- Item-level time stamp data (which are based on keystroke data),
- Contact information,
- Interviewer observations and
- Interviewer demographics.

Item-level time stamp data are based on keystroke data. The interviewer observations include information on the building type, the accessibility of the building and some additional neighbourhood characteristics (e.g. vandalism and public transportation).

Up to the present day, only interviewer observations and interviewer demographics from the first wave of SHARE are released. They were released together with the survey data, since these data also have been collected by asking questions in the course of the survey and therefore were part of the SHARE questionnaire. Subsequently, most of these data were not collected as part of the survey anymore and no paradata has been released, apart from interviewer observations on the interviewer-respondent relation and characteristics of the building[2].

4.1 Collection of Paradata

The collection of paradata is mainly facilitated by using computer-assisted sample management tools and interview instruments. Additionally, interviewer demographics are gathered via excel sheets and delivered by the survey agencies.

[2] These characteristics include information on the building type, on the area where the building is located, as well as information on the number of floors of the building and steps to the entrance.

4.1.1 Process Paradata: CAPI and Sample Management System Data

For sample management, SHARE uses a tailor-made sample management system (SMS), programmed by CentERdata, which is located at the University of Tilburg in the Netherlands. This program is installed on each interviewer's laptop and enables the interviewers to manage their assigned subsample. The success of a cross-national study such as SHARE heavily depends on the way the data is collected in the various countries. Therefore using a harmonised tool for collecting interview data as well as contact data is crucial in order to ensure the comparability of the results. It is for this reason that the sample management system was developed. The SMS tool enables the interviewers to easily register every contact with a household or individual respondent and enter result codes for every contact attempt (e.g. no contact, contact-try again, or refusal). After the contact information is recorded, the SMS manages the start-up of the actual interview (CAPI).

While the interview is conducted, additional paradata is collected by means of tracking keystroke data. Here, every time a key is pressed on the keyboard of the laptop, this is registered and stored by the software in a text file. From these text files, time stamps on item-level can be computed. Additionally, the number of times an item was accessed, back-ups, if a remark was set, and the remark itself are recorded.

4.1.2 Auxiliary Paradata: Interviewer Information and Interviewer Observations

Additionally, two sorts of auxiliary paradata are collected in SHARE. First, interviewer observe characteristics of the sampled household, e.g. if the building is a single house or a multi-story building, if there is evidence for children or persons with disability in the household, or if there is an intercom. The interviewer enters this data partly within the aforementioned SMS or directly into the CAPI system at the very end of the survey interview (in the so-called IV-module). Secondly, information on interviewers' demographics (year of birth, education, gender) and their previous experiences in conducting interviews are obtained from so-called interviewer profiles. The interviewer profiles are Excel-sheets which are filled out by the survey agency since SHARE wave 3. Only during the first wave, the interviewer information was collected in the IV-module of the CAPI interview and had to be entered by the interviewer after every completed interview. However, meanwhile, this procedure has been abolished.

4.2 Data Cleaning and Processing

The preparation of these paradata has been subdivided into several subtasks. First, relevant indicators for the analysis of response behaviour were developed on a theoretical basis. Secondly, so-called generated variable (gv) modules were created that included the previously developed indicators. And finally, the correct linkage of the new modules to the already released survey data was ensured.

When creating the paradata set, particular attention was paid to the format of the paradata set being similar to the format of the already released survey data, since this has several advantages: On the one hand this makes it easier to link the data and on the other hand it makes the data more convenient to use for researchers who are already familiar with the structure of the SHARE data set. For most of the data this includes an aggregation step which reduces the sensibility of the data. The gv-modules were created following the technical procedure of other already existing gv-modules, e.g. the SHARE modules on health, social networks or ISCED-coding of education. All in all, five modules were created, which are based on three different data sources and entail different kinds of paradata. *Table 1* shows details on the underlying data sources of the gv-modules, as well as the indicators derived.

Module	**Data source**	**Linkage ID**	**Indicators**
gv_ks	Keystrokes (time stamps after each item, tracked in Blaise)	mergeid	• Length of interview • Number of items asked • Length for each module • Number of items for each module • Last module in case of breakoff • Length of selected introduction items
gv_ks_xt	Keystrokes for end-of-life interviews (time stamps after each item, tracked in Blaise)	mergeid	• Length of interview • Number of items asked
gv_sms	Sample Management System (contact information on respondent level entered by interviewer)	mergeid hhid	*All indicators on respondent level* • Completed interview in wave 1/2/3/4 • Number of contacts • Number of contact attempts • Contact attempts until first contact • Contact attempts until interview/final refusal • Refusal conversion
gv_smsh	Sample Management System (contact information on household level entered by	hhid	*All indicators on household level* • Household participated in wave 1/2/3/4 • Number of contacts • Number of contact attempts

	interviewer)		• Number of refusals • Contact attempts until first contact • Contact attempts until first interview/final refusal • Final household state • Mode of first contact with household
gv_iwer	Interviewer Profiles (information on interviewers provided by survey agencies)	iwerid	• Year of birth • Gender • Education (ISCED) • Experience as interviewer (in years) • Experience with CAPI (in years) • Participation in SHARE wave 1/2/3

Table 1: Details on gv-modules for paradata
Note: all modules additionally include the variables wave, country, and language.

4.3 Linking Paradata with Survey Data

For ensuring the correct linkage of the different sources, the keystrokes and SMS data run within the same master program as the survey data (CAPI data). Therefore corrections and drops made in the CAPI data are processed on the keystrokes and SMS data in the same way. Minor mismatches between keystrokes, CAPI data and SMS data remain, but bugs have been eliminated and corrections have been carried out to the greatest possible extent.

The interviewer information was collected via excel sheets from all countries in so-called interviewer profiles and needed to be harmonized. For information on the level of education, we applied the International Standard Classification of Education (ISCED-97). For matching the interviewer profiles with the CAPI data, single case cleaning was needed for the interviewer ID (iwerid). At the end of each interview, the respective interviewer is asked to enter her/his ID into the CAPI questionnaire, which is susceptible to typographical errors but the most reliable information for linking the interviewer profiles. Therefore the linkage required extensive data cleaning. In countries, where the interviewer ID entered in the CAPI did not show a similar pattern as the iwerid in the interviewer profiles, interviewer identification relies on using the information on the interviewer provided in the SMS data. *Figure 1* on the following page shows an overview of this procedure, including the different generated modules.

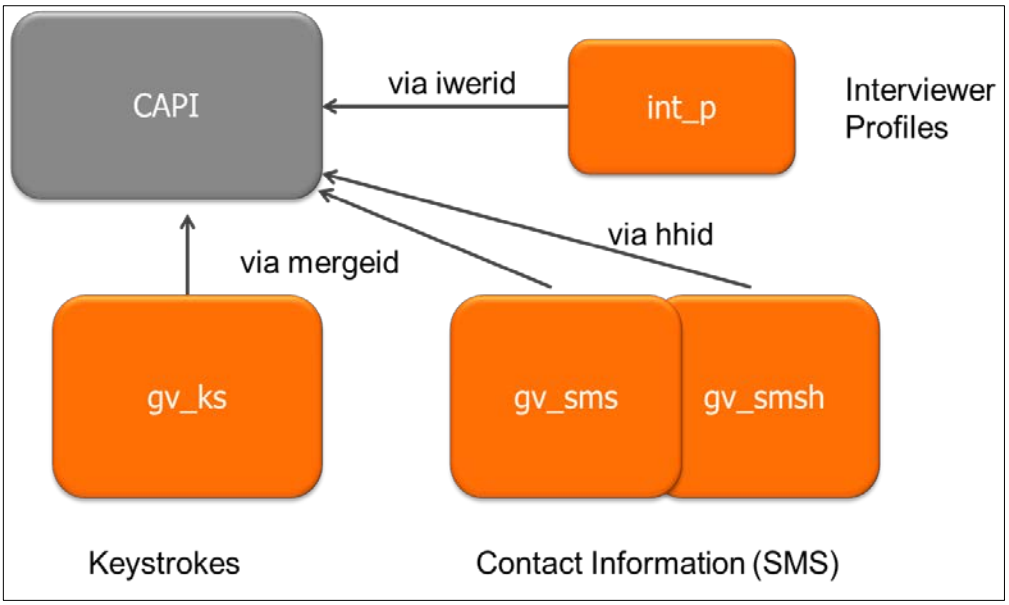

Figure 1: Graphical presentation of linking the paradata sets to the existing CAPI data Abbreviations: iwerid = interviewer ID, hhid = household ID, merge-id = ID of SHARE respondent (scrambled), gv-modules: please see Table 1.

5 Legal and Ethical Issues

Simultaneously to the compilation and linkage of the SHARE paradata set as described above, the extent and nature of ethical issues and legal requirements related to different kinds of paradata have been explored.

As already mentioned in the introduction, legal and ethical consideration with regard to the collection, use and release of paradata are unclear. Only a few researchers have started to address this issue (Kreuter, 2013: 8) and even these authors state that "[e]xisting ethical codes are not very clear on the issue of paradata" (Couper and Singer, 2013: 58). Regarding this, especially the technological developments and the increasing use of computerised systems during data collection amplify (rather than transform) the nature of research ethics and legal requirements relating to research and give rise to specific issues which are not covered in existing ethics codes.

For survey researchers two key ethical principles (cf. Singer, 2008: 85; Couper and Singer, 2013: 57) are

- to prevent respondents from harm and
- to assure the autonomy of the respondents.[3]

[3] Besides the key issues of informed consent and confidentiality protection, additionally a third principle of 'justice' is advanced for the conduct of research involving human subjects. However, this principle, which

This means in practice that researchers have to ***ensure the confidentiality*** of the data they collect from their respondents and that they have to ***obtain informed consent***[4] of their respondents prior to data collection. However, 'Codes of Ethics' for survey researchers do not constitute rules. According to Denscombe,

> "[t]he point is not that each principle should be *followed*, but that it should be taken into account and *considered.* Each principle provides a starting point, a baseline against which to compare the actual position adopted by the researcher. If circumstances arise where the researcher feels that he or she is not able to be bound by a specific principle it becomes necessary to weigh the pros and cons of the situation and to arrive at a decision about whether it is legitimate to 'relax the rules' on this occasion. To do so does not automatically condemn the research as 'unethical' but it does warrant some explanation. *The principle should be acknowledged*." (Denscombe, 2002: 176)

Since Paradata by definition is always closely related to the process of survey production, whether recorded as a by-product in the course of conducting a survey or obtained separately from external sources or with a specifically targeted effort, the same key ethical principles have to be considered when collecting, processing, re-using and disseminating paradata.

5.1 Data Collection and Usage of Paradata – Informed Consent

The issue of ***obtaining informed consent*** of the respondents in survey research mainly relates to data collection and data usage. Usually consent – whenever it has to be obtained, whether in a written or a verbal form – has to be obtained prior to data collection. According to Singer, "[i]nformed consent requires (a) providing enough information about potential benefits and risks of harm to permit subjects to make informed participation decisions; (b) assuring that the information is understood: and (c) creating an environment that is free from undue influence and coercion" (Singer, 2008: 85).[5]

When looking at paradata, first it has to be noted that there are several types of paradata that can be collected and that there are also different ways of recording these data. The different ways of collecting paradata as well as the amount and types of paradata that can be collected are heavily depending on the way in which a survey is administered, i.e. the context in which paradata are collected (e.g. web surveys, CAPI, CATI, mixed-mode etc.).

---

aims at a fair balance between the subjects who bear the burden of research and those who benefit from it, is more important to biomedical research. (cf. Couper and Singer, 2013: 57; Singer, 2008: 80).

[4] It is noted that "obtaining respondents' informed consent […] has nothing to do with protecting subjects from harm, and everything to do with assuring that they are treated as autonomous individuals with the right to make informed, voluntary decisions about participation". (Couper and Singer, 2013: 57)

[5] "In addition, (d) research organisations ordinarily need some evidence that subjects have, in fact, been adequately informed and have agreed to participate." (Singer, 2008: 85)

Therefore, the legal and ethical issues that are connected to the collection and use of paradata require a nuanced approach.

Especially with regard to the issue of obtaining informed consent, it becomes obvious that new technologies allowing unprecedented levels of data collection, data collation and data dissemination, amplify ethical challenges and give rise to specific issues which are not covered in existing ethics codes.

The rapid growth regarding the collection of paradata and the increasing use of these data is accompanied by the development of new applications and monitoring systems. At this, particularly the *intended use* of the paradata appears to be crucial. Since process paradata, defined as a by-product of survey production, are unavoidably collected in the process of survey production the only relevant question is whether respondents would consent to their 'use' (cf. Couper and Singer, 2013: 65). Regarding this, Couper and Singer state that:

> "While most […] studies focus on improving the quality of research procedures and, particularly, the questionnaire, paradata are increasingly being used to enhance other information provided by respondents – that is, turning from purely methodological research to more substantive research. There is no consensus on whether, or under what conditions, respondents should be informed that paradata are being collected and may be used. Arguably, they ought to be informed if researchers plan to use such data in conjunction with other information provided by respondents in order to make inferences about individuals. In other words, as the paradata (information about the process) are turned into data (information about respondents), informed consent issues may arise." (Couper and Singer, 2013: 57)

From an ethical perspective, the capturing of process paradata, can be understood as "nothing more than collecting information about the process of completing a survey that is already covered by the informed consent statement for the survey itself" (ibid.: 59). Since paradata by definition is data about the process of survey production, which does not capture respondents' behaviour outside the survey, it can be argued that no additional consent is needed in the case of process paradata. However, the questions of whether, how and to what extent participants should be informed about the capture and the use of paradata remain, if it is assumed that

> "respondents are not aware that such additional information is being collected, do not have a reasonable expectation of such capture and use, and, if they were aware of it, might change their behavior or decide not to participate in the survey. Under these circumstances, difficult questions arise about how best to provide information about the collection of paradata while at the same time maintaining respondent cooperation with the survey" (Couper and Singer, 2013: 59).

Obviously, auxiliary paradata is not 'unavoidably' collected in the course of a survey, as these data are obtained separately from external sources or collected with a specifically targeted effort. Here, the question whether respondents would consent to their collection may still be of relevance. However, since these data are collected in order to enhance the information on the survey production process and, especially, if auxiliary paradata do not constitute information on the respondent, this question also may not arise.

Accordingly, the question of whether or not additional informed consent[6] is needed in the case of auxiliary paradata collection is highly dependent on (a) the question whether the data can be classified as information on the respondent and on (b) the question whether it is being used to enhance the information provided by respondents in the course of the survey. Here, on the one hand, it can be argued that respondents should not only be informed about the capture of paradata but also additional consent of them has to be obtained, if any of these conditions apply to the collected auxiliary paradata.

On the other hand, it can be argued, if none of the two conditions applies to auxiliary paradata, neither obtaining consent nor providing information to the respondents are needed.

5.2 Data Processing and Data Release – Confidentiality Issues

While the issue of obtaining informed consent of the respondents primarily relates to data collection and data usage (even though, of course, the participants also have to be informed about the way the data is processed and how it will be released), ***ensuring the confidentiality*** of the respondents' data becomes crucial in relation to data processing and data release.

Ensuring confidentiality of data collected in the course of population based surveys is of crucial importance since "most serious risks of harm to which participants in social research are exposed are breaches of confidentiality, and the consequences that may follow from such breaches" (Singer, 2008: 90).[7]

Since in the context of most surveys some sensitive or confidential information is collected that might lead to negative economic, social, psychological consequences (such as the loss of employment, the loss of reputation, stigmatisation and discrimination or even criminal penalties), if revealed to unauthorised others, it is crucial not to disclose the identities of the participants. Therefore, particular importance has to be placed on the compliance with European and national/regional data protection law as well as on the safeguarding of

[6] It is supposed, that the respondents have to agree respectively have agreed to participate in the survey.

[7] Besides, it is noted that breaches of confidentiality do not only pose a risk of harm to survey participants, but also to the 'survey enterprise' itself. (cf. Singer, 2008: 91)

sensitive data and confidential information. This, of course, also has to be taken into account with regard to the processing and the release of paradata and all measures necessary to ensure data privacy have to be taken.

In general, two kinds of variables, which could be used to disclose respondents' identities, can be identified. "[A] person's identity can be disclosed from:

a) direct identifiers such as names, addresses, postcode information, telephone numbers[, ID numbers,] or pictures

b) indirect identifiers which, when linked with other publicly available information sources, could identify someone, e.g. information on workplace, occupation or exceptional values of characteristics like salary or age"
(UK Data Archive, accessed 27/06/2013)

If paradata is collected in the context of a survey and the use of these data is intended, the paradata sets should be checked for both direct identifiers and indirect identifiers prior to the use or release of the data. Furthermore, the data environment has to be considered. Especially, if linking paradata to the survey data set is intended, additional attention has to be paid to relational data, i.e. to variables in the paradata set and the survey data set that might lead to the disclosure of identities when being connected.

If certain paradata collected in the context of a survey (e.g. contact protocols or interviewer observations) include direct identifiers, indirect identifiers or relational data that might lead to a disclosure of the identities of respondents, appropriate measures have to be taken in order to ensure the confidentiality of the data (unless the respondent explicitly has given consent to use/release these data).[8]

While direct identifiers, which are often collected in the course of survey administration, can be removed from the data easily, since these usually do not constitute information that is needed in the context of methodological or scientific research (cf. UK Data Archive), removing indirect identifiers or relational data that could lead to disclosure of identities, might pose a more challenging task for survey researchers. Here, anonymising data (e.g. removing or aggregating variables[9]) or pseudonymising data (i.e. disguising identities, whilst retaining the possibility to backtrack to the individual under predefined circumstances[10])

[8] I.e., all technical and organisational measures as laid down the relevant national (or regional) legislation of the current member states of the European Union that implement the European Data Protection Directive (95/46/EC) have to be taken.

[9] Besides removing or aggregating variables, other techniques for handling risk disclosure might be applied in order to ensure the confidentiality of data. A number of commonly used options when dealing with variables, which might act as indirect identifiers, can be found on CESSDA's website.

[10] This, for instance, is necessary in panel studies, such as SHARE, that need to re-contact the participants of previous waves.

might result in a loss of data usability.[11] Even though anonymisation and pseudonymisation are central security measures to ensure data confidentiality, alternative measures, such as access restrictions, maintaining the usability of the data should be considered when preparing paradata including sensitive information for use and re-use. Concerning this matter, CESSDA generally states:

> "Anonymisation is often the first approach considered by most researchers, but this should not be considered in isolation. Sensitive and confidential data may also be safeguarded effectively through access and usage restrictions employed in certain circumstances and if deposited in a formal archive" (CESSDA, accessed 27/06/2013).

So-called 'public-use files', which can be released to the scientific community and even the entire public may only contain anonymised data. Inevitably, such data collections are limited with regard to their usability for scientific and methodological research since some information has to be removed from them and some of the data contained has to be adjusted through data-masking procedures (cf. CESSDA, accessed 27/06/2013). However, in most[12] countries more sensitive data may be analysed in data enclaves, since in such a restricted (data) environment, the opportunities for de-anonymisation are reduced to a substantial extent.

With regard to the release of confidential paradata it might therefore be worthwhile to consider other levels of access – such as 'on-site use' (i.e. analyses of data in separate secure workplaces for guest researchers) or 'remote data access' (indirect access to confidential microdata)[13] – which allow making more sensitive and less anonymous versions of the data available for scientific analyses to vetted users.

6 Concluding Remarks

In several experiments on the effects of asking consent for paradata collection on web survey participation, Couper and Singer have found that "the concept of paradata is inherently difficult to grasp and is unfamiliar to virtually all respondents [and that t]he potential uses that might be made of such data are equally mysterious [to respondents]"

[11] "Anonymisation is a very valuable tool, allowing sensitive data to be shared whilst preserving privacy. Of course, anonymising data makes them less useful than accurate, fine-grained data." (UKAN – UK Anonymisation Network, accessed 27/06/2013)

[12] This depends on the concrete implementation of the Data Protection Directive (95/46/EC) in the respective EU member state. While the Directive 95/46/EC includes a minimum set of provisions to be implemented by the member states, the member states are free to 'increase' the level of data protection for their country.

[13] Remote Data Access (RDA) allows researchers to submit their own computer programs to research data centres (RDCs). At the RDCs, these will be run on the confidential microdata sets. Subsequently, after having been scrutinized for confidentiality, the results are returned to the researchers.

(Couper and Singer, 2013: 65).[14] On the one hand, this indicates that asking for consent might, in the first place, raise awareness about process paradata that is unavoidably being collected in the process of survey production, which again might change the behaviour of respondents or even result in them deciding not to participate.[15] On the other hand, this finding points to the issue of how participants can be informed about paradata collection and how much detail should be provided to them, while at the same time avoiding a decrease of participation rates – which might turn out to be a challenging task for survey researchers.

When investigating the ethical aspects and the legal requirements related to paradata, it becomes clear that on a general level this *question of whether, how and to what extent participants should be informed about the use of paradata* is the major contentious issue with regard to paradata. Regarding this issue, Couper and Singer state that "[t]he question of whether the use of paradata […] rises to a level needing explicit mention to respondents remains an open one" (Couper and Singer, 2013: 66). Even though Couper and Singer focus on web surveys in their paper, this finding equally applies to the collection of paradata in survey research using CATI or CAPI techniques. Moreover, it also equally applies to process paradata and auxiliary paradata whenever information about the collection of these data is to be given to the respondents.

However, from an ethical perspective, this is not the only issue that remains; additionally, in some cases of collecting auxiliary paradata the *question if additional informed consent should be obtained* remains open.

Taking into consideration that there are many different kinds of paradata that can be collected, depending on the survey mode and the technical system in place, and that the various kinds of paradata (such as keystroke data or contact protocols, as collected in the context of SHARE) only can be used for certain kinds of analyses, these questions might need to be answered on a case-by-case basis, taking into account the specific kind of paradata, the concrete context in which these data are collected and how they are and may be used.

For example, if interviewer demographic characteristics are collected in order to enhance the information on the survey production process, very specific issues, exceeding the ones described in this document so far, have to be taken into account. When collecting and using these information the interviewers themselves become data subjects and have to be

[14] According to Couper and Singer (2013), many respondents, for example, confuse paradata collection with behavioural tracking (cf. ibid.: 61).

[15] Actually, this also can be identified as the major finding of Couper and Singer (2013): "In all three experiments, any mention about capture of paradata lowers stated willingness to participate in a hypothetical survey" (ibid.: 57). However, since the experiments only considered web surveys – which obviously in the experiments have been associated with general threats to privacy occurring on the internet (e.g. browser-related, IP-related, behavioural tracking) – this finding may not apply with regard to computer-assisted face-to-face or telephone survey data collection.

considered in this role, including specific needs and rights, as well. Now, researchers do not only have to ensure the confidentiality of the data collected in the survey and obtain informed consent of their respondents, but also have to consider these issues with regard to the interviewer. Furthermore, besides ethical issues and data protection requirements, in some cases (depending on the information included in the interviewer profiles and on the way in which these are obtained) national employment legislation has to be considered.

This example shows that there are no general rules on how to deal with the variety of 'micro-level data about the process of survey production' (i.e. paradata) that are collected as part of or in addition to web-based, computer-assisted face-to-face or telephone surveys or mixed-mode surveys using various applications, instruments and monitoring systems. Denscombe's assessment of ethics principles corresponds to this. When emphasising that ethics principles do not constitute rules that have to be followed, but rather are recommendations that should be considered, taking into account the pros and cons of the specific situation (cf. ibid., 2002: 176), he acknowledges that specific situations may require specific solutions.

While ethics guidelines and frameworks generally set out different positions, the law generally sets out what can and cannot be done. With regard to the collection and use of paradata, especially legal provisions regarding issues of confidentiality and data protection are of relevance.

However, when taking into account the current European data protection legislation, first, it is to be noted that the legislative regime which impacts upon the governance of the process of survey research, including ethics, from study initiation to data dissemination, is marked by fragmentation and uncertainty. Currently, the legal basis of this regime is the "Data Protection Directive" (95/46/EC) and the implementation of its provisions in form of national/regional[16] data protection laws. However, since the provisions of the Directive have been implemented in different ways in the member states, differences in the level of data protection, both in paper and practice, exist.

Besides this fragmentation of laws, there is also uncertainty with regard to on-going negotiations relating to the EC Proposal of a "General Data Protection Regulation" that will affect the data protection regime currently in place. Whilst this new Regulation aims at reducing the existing fragmentation and harmonising legislation and legislative practice throughout Europe, at this point in time, neither the concrete provisions that finally will be included in the Regulation nor its possible implications with regard to survey data collection are clear.[17]

[16] For example, in Germany in addition to the Federal Data Protection Act ("Bundesdatenschutzgesetz"), each German state ("Bundesland") has its own data protection law.

[17] Cf. Amendments to the Proposal of a General Data Protection Regulation (GDPR).

Currently, national and in some cases regional differences inevitably lead to difficulties when, for example, trying to assess, under which circumstances informed consent to the collection and use of paradata has to be obtained or to what extent participants should be informed about the use of paradata. And, not even with regard to much more straightforward matters, such as the linkage of survey data with administrative record data, uniform procedures exist: while, for instance, in Denmark currently no consent has to be obtained when linking survey data and administrative record data, in Germany written informed consent is obligatory.

With regard to data access and usage restrictions (e.g. via 'on-site use' or 'remote data access') concerning the release and re-use of sensitive or confidential information, including different degrees of anonymisation, similarly, differences in the level of data protection may be experienced between different EU member states. Such differences also have to be taken into consideration with regard to the processing and the release of paradata, especially when paradata are to be shared in a cross-country usage scenario.

Furthermore, from a legal perspective (especially if administering web surveys), recent EU online privacy legislation and associated national laws have to be taken into account, such as the "new e-Privacy Directive" (2009/136/EC) (cf. ESOMAR, 2012), which finally may also affect the collection and use of paradata not only in relation to web surveys, but also with regard to survey research in general. "While the intent of [regulations of this kind] is to limit online behavioral tracking, they may encompass a number of more benign activities such as paradata capture in surveys" (Couper and Singer, 2013: 66).

Similarly to ethical issues, legal requirements related to the collection and use of paradata also require a nuanced approach. Here, not only national and regional differences in the level of data protection have to be taken into account as long as there is no EU-wide data protection regulation in place; also the nature of the specific kinds of paradata and the mode of collection have to be considered. And, even when the General Data Protection Regulation enters into force, still questions such as whether consent has to be obtained for specific kinds of paradata or whether, how and to what extent participants should be informed about the use of certain kinds of paradata might remain open.

Therefore, besides monitoring closely the legislative processes, it is important to further investigate systematically the ethical issues related to different types and kinds of paradata and to try to answer ethical questions that may arise on a case-by-case basis – taking into account the specific paradata that are concerned (including their ways of collection as well as actual and potential use cases), the survey mode which is applied and the data environment in which paradata collection, processing, usage and release are taking place.

7 References

CESSDA (n.d.). Confidentiality. In: Council of European Social Science Data Archives - CESSDA. Retrieved June 27, 2013, from http://www.cessda.org/sharing/rights/3/.

Couper, M.P. (1998). "Measuring survey quality in a CASIC environment." Proceedings of the Section on Survey Research Methods Section, American Statistical Association, pp. 41-49.

Couper, M.P. and Lyberg, L. (2005). "The Use of Paradata in Survey Research." Proceedings of the 55th Session of the International Statistical Institute.

Couper, M.P. and Singer, E. (2013). "Informed Consent for Web Paradata Use." Survey Research Methods 7(1), pp. 57-67.

Denscombe, M. (2002). "Ground Rules for Good Research: A 10 point guide for social researchers." Open University Press, Buckingham.

ESOMAR (2012). "ESOMAR Practical Guide on Cookies." Retrieved June 27, 2013, from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Practical-Guide-on-Cookies_July-2012.pdf.

Groves, R.M. and Heeringa S.G. (2006). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." Journal of the Royal Statistical Society 169(3), pp. 439-457.

Heerwegh, D. (2002). "Describing response behavior in websurveys using client side paradata." Paper presented at the International Workshop on Websurveys, pp. 17-19 October 2002, Mannheim, Germany.

Kennickell, A., Mulrow, E. and Scheuren, F. (2009). "Paradata of Process Modeling for Inference." Modernisation of Statistics Production Conference, Stockholm, Sweden.

Kreuter, F. (2013). "Improving Surveys with Paradata: Analytic Uses of Process Information." Wiley.

Kreuter, F. and Casas-Cordero, C. (2010): "Paradata." RatSWD, Working Paper Series No. 136.

Singer, E. (2008). "Ethical Issues in Surveys." In: De Leeuw, Edith D. / Hox, Joop J. / Dillman, Don A. (eds.). International Handbook of Survey Methodology. Psychology Press, Taylor & Francis, New York, pp. 78-96.

UKAN (n.d.). "What is Anonymisation?." In: UK Anonymisation Network. Retrieved June 27, 2013, from http://www.ukanon.net/key-information/.

UK Data Archive (n.d.). "Anonymisation." In: Create and Manage Data. Retrieved June 27, 2013, from http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation.

8 Acronyms and Abbreviations

CAPI – Computer-assisted personal interviewing

CATI – Computer-assisted telephone interviewing

CESSDA – Council of European Social Science Data Archives

DASISH – Data Service Infrastructure for the Social Sciences and Humanities

DoW – Description of Work, Annex 1 to the Grant Agreement of the DASISH project

EC – European Commission

EU – European Union

ESOMAR – European Society for Opinion and Market Research

GDPR – General Data Protection Regulation

ISCED – International Standard Classification of Education

RDC – Research Data Centre

SHARE – The Survey of Health, Ageing and Retirement in Europe

SMS – Sample management system

RDA – Remote Data Access

SSH – Social Sciences and Humanities

WP(#) – Work Package(Number)