Introduction to the Management of Social Survey Data

Jindřich Krejčí


Data Management
coding
archiving
planning
international
documentation
Social Survey
Research
values
formats
comparative
backups
cleaning
Copyright
process
Data Life Cycle
Open Access
preservation
variables
Methods
data file
strukture

Introduction to the Management of Social Survey Data

Jindřich Krejčí

ČSDA

SOÚ

Institute of Sociology CAS

Institute of Sociology CAS

Prague 2014

Author:
Mgr. Jindřich Krejčí, Ph.D., jindrich.krejci@soc.cas.cz

Reviewers:
Mgr. Jiří Remr, Ph.D.
PhDr. Mgr. František Kalvas, Ph.D.

This paper was prepared with the support of the 'CESSDA project: Construction of the Czech node of CESSDA-ERIC and its integration into this large-scale pan-European research infrastructure' which aims to provide data services for socio-economic research financed by the Ministry of Education, Youth and Sports under reg. no. LM2010006. The author would like to thank Jiří Remr and František Kalvas for their reviews and very useful recommendations and Jan Morávek for his help with the English version of the paper.

Keywords:
data management, social survey data, computer data file, data sharing, open access

ISBN 978-80-7330-252-8

CONTENTS

INTRODUCTION

Although the construction of databases, data processing, documentation and other operations with data may have a fundamental influence on research results, the management of data rarely receives attention as a separate subject in the methodology and practice of survey research. More often it gains attention only in the form of individual and isolated tasks during different stages of the research process. This, however, might pose a problem, because there are ample good reasons for paying attention to data management on a systematic and ongoing basis.

The quality of data management affects the quality of data and importantly determines the adequacy of research findings. This also applies at a more general level, namely to the building of a good environment for research work. Systematic data management is, above all, a necessary condition for preventing errors and false findings, but it can also save a lot of time and make research work both clearer and easier. An abundance of stories about ridiculous mistakes caused by analysing incorrect data circulate among researchers; many analyses have failed because of flawed datasets; and numerous researchers have spent long hours disentangling poorly documented data. Much research information has got lost due to disorder in datasets and a lack of effort in their documentation.

Higher demands are placed on data management, compared to independent research work, especially when we plan the long-term preservation and sharing of data between different research teams. In recent years, the requirements of archiving and providing open access to social science data for the purpose of secondary data analysis have become highly important parts of scientific work. Consequently, demands on professional data management have also risen. Moreover, such higher demands are not an end in itself; they arise directly from new developments in empirical social research and methodology.

Above all, there has been long and rapid growth of the volume of data that social research utilises. The reasons include not only the growing production of survey data but also rapid digitisation in general. At the same time, newly developed and implemented information technologies have given rise to new methods of utilizing, transmitting and sharing data. The character of research work these days is also strongly affected by such forms of collaboration that rely on the sharing of datasets among research teams. For example, these practices have become crucial to the development of international comparative research or to the building of an analytic base for evidence-based decision making in public policy. The volume and availability of data sources and new types of data represent an important factor in the development of analytic methods, and, in turn, the introduction of novel methods places additional demands on data production and processing.

Naturally, higher volumes of data require a more systematic approach to data production and processing. These days, the management of social science data sources is, in many aspects, similar to the situation in many disciplines of scientific and engineering research where large volumes of digital information must be processed. Alongside applying information technologies, some working methods and standards originating in the information sciences have also been adopted. The abundance of available research data has also increased the relevance of secondary analysis. However, for these reasons, care should be taken in the production of data and it should be documented to ensure that people outside the researchers' team understand the data; furthermore, the implementation of research projects, whether or not they produce original data, should systematically rely on a vast array of available data sources.

In the academic sector, efficiency in the sharing of digital data is provided by social science digital data archives, which have thus become an important part of the social research infrastructure. Besides data services, they also provide a background for international co-operation, the organisation of comparative research and methodological developments. In the Czech Republic, this kind of services is provided by the Czech Social Science Data Archive (CSDA)[1] at the Institute of Sociology of the Academy of Sciences. The CSDA operates within the framework of the large scale pan-European research data infrastructure of the Consortium of European Social Science Data Archives (CESSDA)[2] and helps integrate Czech data sources and empirical social research into international science.

The CSDA is not just a library of scientific data. Its mission also includes organising Czech participation in international survey research programmes, methodological research, and the general promotion of secondary data analysis. However, secondary analysis also adds significant problems to research efforts. Survey data are not readily understandable by researchers outside the original research team. Published social research papers are often based on sophisticated analysis, but their authors sometimes pay only little attention to understanding the meaning and quality of the data analysed, which they have downloaded from somewhere on the Internet. Thus, not only the dissemination of data, but also the promotion of quality data production, documentation and responsible data processing and analysis are among the priorities of data archives.

In order to respond to the challenges of current developments, namely the fact that data management is growing in importance but does not always receive adequate attention in research projects, the Czech Social Science Data Archive (CSDA) initiated the writing of two books about social science data in the Czech Republic, and the sources and work with such data. The first one [Krejčí, Leontiyeva 2012] is in the Czech language and is devoted to data management, measurement issues and a survey of available data on Czech society; the second one is in English and concentrates on political science data and analysis [Lyons 2012]. In this framework and with the aim of providing more comprehensive information for both Czech and international researchers, the present paper should serve as an introductory guide to data management for English-speaking users of the CSDA's data services. Moreover, we believe that the information provided herein may be useful for all empirically oriented social science researchers

and students, whether or not they are the CSDA's clients. The paper was inspired also by manuals, recommendations and methodological texts published by various other data organisations, in particular the online guide 'Create & Manage Data' by the UK Data Archive [UKDA 2010], the 'Guide to Social Science Data Preparation and Archiving' by the US-based international consortium ICPSR [ICPSR 2012] and the online presentation 'Sharing Data' by CESSDA [2009].

The paper opens with a discussion of the conceptual background to data management in contemporary social research, which is shaped by efforts to implement open access policies and a cyclical view of the life of data. It then briefly reviews the principles of data management in the preparation stage of the research process. Finally, it guides the reader through the different areas of data management during the research process.

[1] See *http://archiv.soc.cas.cz* (visited 12 September 2013).

[2] Since 1970 the CESSDA has served as an informal umbrella organisation for European national data archives. In 2013 the CESSDA established a consortium as a legal entity to build a new international European research infrastructure based on interconnection of current national data services. The main office is located in Bergen, Norway. A new system of pan-European data services is under construction. See *http://www.cessda.org* (visited 12 September 2013).

1. THE FOUNDATIONS OF CURRENT METHODS OF DATA MANAGEMENT

1.1. A POLICY OF OPEN ACCESS

The exchange of scientific information and results is crucial for the development of research today. This is equally true for the sharing of social science data. Data availability and the possibility to utilise and combine various publicly available databases contribute to the work of numerous research projects. Data sharing practices also provide an important basis for international comparisons and the study of developments over time. Furthermore, open access to data sources makes it possible to utilise research data for instruction at universities. Knowledge of prior projects' data outcomes facilitates the formulation of new research projects and reduces the need to repeatedly conduct empirical surveys. Data availability increases the verifiability of results and the transparency of scientific research.

Data are often produced at considerable expense for the public resources. However, the general public receives value in return not immediately after the construction of the database, but only when it is analysed and new knowledge is generated. At the same time, the informational value of data usually goes beyond an individual project's research intent. Therefore, it is logical to ask researchers who are recipients of public funding to make further use of their databases possible. In other words, when they are done utilising the dataset within their own research project, they should make it available for secondary analysis whenever this is not precluded by the nature of the data or any other specific circumstance.

The exchange of data among different research teams can be done on a commercial or a reciprocal basis. In the academic environment of universities and public or governmental institutions, and when research receives public funding for the purposes of creating publicly available results and knowledge, it is possible to organise the wide and effective sharing of research data on the basis of 'open access' data repositories. In the social sciences this practice has become a general standard, whenever open access is not prevented by the nature of the data or specific circumstances of its production.[3] According to OECD--defined principles, openness in this context means 'access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.' [OECD 2007: 15] Current practices of organisation of social science research based on data sharing are also gradually institutionalised on the level of international and national science policies. For example, in 2004, OECD members and several other countries signed the 'Declaration on Access to Research Data from Public Funding' [OECD 2004], which commits parties to implementing a set of basic goals of access to research data. Subsequently, the OECD [2007] codified a set of basic principles[4] underlying contemporary data policies. These principles have also been adopted by the European Union, among others.[5]

Specific measures encouraging researchers to share data have been implemented by many EU member states and other developed countries, primarily at the level of public funding agencies [see Ruusalepp 2008]. Numerous grant agencies apply highly precise and strict rules of access to data. For example, the US National Science Foundation stipulates the requirement of data sharing in its General Grant Conditions [NSF 2006: 27, par. 38] and simultaneously defines detailed sets of specific requirements for different scientific disciplines and programmes.[6] Seven scientific councils covering different disciplines in the UK have a shared set of general principles and, based on this set, each of them defines its own data policy for its area of funding.[7] For example, the Economic and Social Research Council (ESRC) requires a specific procedure for making data available within defined structures. Researchers will not receive final grant appropriation unless they provide evidence that their data have been published [ESRC 2010: 4]. The Czech Republic has also agreed to abide by these principles of open access. Czech funding agencies can therefore be expected to implement adequate measures in the future as well.

The conditions for data sharing are determined not only by the requirements of governments and funding agencies, but also, more importantly, by the existence of an adequate research infrastructure for depositing and distributing data. Sometimes open access to data in social research can be provided using the resources of individual projects. However, this solution tends to be efficient only for longitudinal surveying programmes which create cumulative databases and are capable of ensuring long--term access to them (e.g. the pan-European projects of the ESS and SHARE, the German ALLBUS or the American GSS)[8] or for other extensive collections of data with permanent data management (e.g. comprehensive collections of data for certain specific research topics). In other cases, social data is normally made available by means of centralised national or international data archives.

[3] Thus, for example, open access is not required for research projects funded by means of public procurement or conducted for commercial purposes. Access can be further restricted for the reasons of specific ownership arrangements, protection of personal data, protection of intellectual property, national security, trade secret or interference with legal processes. Further legitimate reasons include technical barriers to open access, international commitments etc.

[4] The 13 principles outlined by the OECD are as follows: Openness, Flexibility (in terms of the development of IT, research systems, legal systems etc.), Transparency (availability of relevant documentation), Legal Conformity (legal and legitimate protection of rights and interests, including personal data, trade secrets, national security etc.), Protection of Intellectual Property, Formal Responsibility (formal arrangements for ensuring access), Professionalism (conformity with professional standards), Interoperability (technological compatibility and international standardisation), Quality (data and metadata quality, verifiability), Security (guaranteeing the integrity of datasets, protection of information against loss), Efficiency (efficiency of data access, utilisation, management etc.), Accountability (data evaluation, monitoring of utilisation), and Sustainability (long-term preservation and access). The principles only apply to data produced with public funding and for the purposes of creating publicly available results and evidence.

[5] The European Commission formulated an action plan in its Communication on Scientific Information in the Digital Age: Access, Dissemination and Preservation [European Commission 2007]. Tasks for the member states and the European Commission are summarised in the European Council's Conclusions on Scientific Information in the Digital Age: Access, Dissemination and Preservation [Council of the European Union 2007], which includes a requirement to 'reinforce national strategies and structures for access to and preservation and dissemination of scientific information' in EU member states. Open access to scientific data has become an integral part of EU agendas on research and innovations and the knowledge-based economy, including a new support program Horizon 2020. In 2012, the European Commission outlined measures to improve access to scientific information in a Communication [Council of the European Union 2012a] and a Recommendation [Council of the European Union 2012b] to Member States (visited 12 September 2013).

[6] See, e.g., NSF Data Archiving Policy: *http://www.nsf.gov/sbe/ses/common/archive.jsp* (visited 12 September 2013).

[7] UK data policies portal: *http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies* (visited 12 September 2013).

[8] The European Social Research (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE), the German General Social Survey (ALLBUS), the General Social Research (GSS) in the United States.

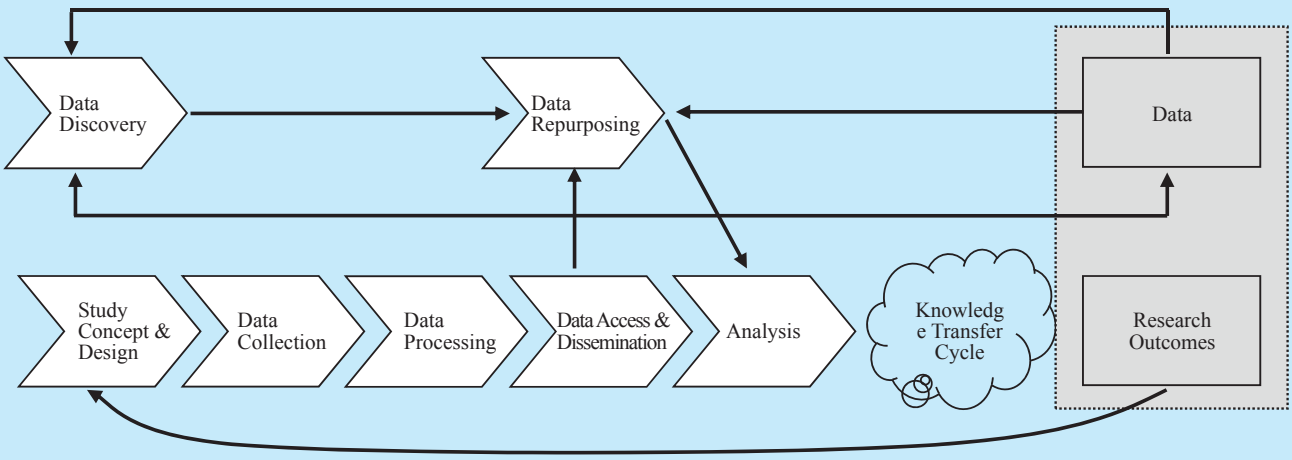
1.2. THE DATA LIFE CYCLE

The requirements of data sharing have changed the functions of data management. When a dataset is being produced, one must count on archiving and publicising it without knowing by whom and for what purposes the data are going to be used. This naturally affects the preparation of datasets, the safeguards taken during that process and the requirements of data documentation. Moreover, not only individual tasks but also the entire concept of data management is affected.

Like in other disciplines, research in the social sciences takes the form of a cycle where the results of one research study feed back into the research process as background for other research studies. In an environment characterised by open access to data, secondary analysis of research data plays an important role in this cycle [e.g. Humprey 2006, Green and Gutmann 2007]. This gives data a new function in the dissemination and reproduction of knowledge and this function must be reflected in the ways data are managed. Thus, an important basis for contemporary approaches to social data management is provided by the concepts of the 'data life cycle'.

In such concepts, the management of digital information is incorporated into a cyclical system of creating scientific knowledge. For example, Charles Humprey's [2006] life cycle model is depicted in Figure 1.

FIGURE 1: THE LIFE CYCLE MODEL OF RESEARCH KNOWLEDGE CREATION



Source: Humprey [2006].

Data life cycle management represents a comprehensive set of methods. It relies on a model of data utilisation in the course of its individual stages, with their different goals, functions and actors. Relations between the model's elements are determined by the life cycle of scientific knowledge. Such approaches are applied, for example, in the UK Data Archive's data management methods guidelines [UKDA 2011, UKDA 2010, Eynden, Corti et al. 2011, Eynden, Bishop et al. 2010] or by the ICPSR consortium in the United States [ICPSR 2012]. This concept also figures in the international data documentation standard 'DDI Life Cycle – DDI 3.0', which allows data descriptions to be built on a continual basis and with respect to different phases of the data life cycle [see, e.g., Iverson 2009].

Table 1 depicts the different stages of data management in the model utilised by the ICPSR for depositors and users of its data archive. In it, a research project begins with a review of existing data, and data management permeates the entire research process.

These approaches to the data life cycle also influence the concept of data management presented in this paper. The above-presented Humprey Model provides a general idea of the role of data management in the research process. A similar logic is followed by the phasing of data management in the ICPSR manual. In our case, however, the focus is on describing the different main roles of data management, some of which are relevant to several stages of the research process.

TABLE 1: PHASES OF SOCIAL SCIENCE DATA MANAGEMENT ACCORDING TO THE ICPSR

Phase 1	Proposal Planning and Writing	• review existing datasets • determine the need for a new dataset • describe special archiving challenges, e.g. informed consent and confidentiality • identify potential users • describe costs related to archiving
Phase 2	Project Start-Up and Data Management	• create a data management plan • make decisions about documentation form and content • conduct pre-tests and pilots of materials and methods
Phase 3	Data Collection and File Creation	• follow best practices, • data entry and digitisation, dataset integrity, variable names • and labels, groups, coding, missing data, • document data: include all relevant documentation elements • document constructed variables
Phase 4	Data Analysis	• control dataset editions and versions control • set up appropriate file structures • back up data
Phase 5	Preparing Data for Sharing with Others	• address disclosure risk limitations • determine file formats to deposit • contact archive for advice
Phase 6	Depositing Data	• complete relevant forms • comply with dissemination standards and formats
Phase 7	After Deposit – Archival Activities	• collection evaluation • storage, migration and metadata creation • additional confidentiality review • possible preparation for on-line analysis and data • enhancement • preservation of data • support for data users

Source: ICPSR [2009, see also 2012].

2. DATA AND RESEARCH DESIGN

There are at least five good reasons to pay attention to data management from the very beginning of the planning of a proposal:

1. Available existing databases can often be utilised as sources of empirical data, replacing or supplementing our own data collections. Moreover, a combination of several data sources may make time or cross-national comparisons possible.

2. Such databases and especially their documentation can also be used in preparing the research instruments and the design of a new survey, in replicating or verifying some procedures, or as inspiration for some tasks of research planning.

3. Data processing is related to a series of formal and legal conditions that must be met in order to implement a research study, especially in the fields of personal data disclosure and copyright.

4. Systematic, well-prepared dataset management will improve data quality and thus the reliability of research results.

5. Data operations are not for free and their costs should be considered in budgeting.

2.1. REVIEWING DATA SOURCES

For the above-mentioned reasons every proposal for empirical research should start not only by reviewing the relevant literature but also by carefully reviewing available data sources. This is equally important for studies that are based on their own data collection, because besides data one can also utilise the accompanying information about methodologies, procedures and research instruments from prior research. For that purpose, researchers planning a new proposal should always browse data archives and other scientific data sources for the following information:

- the availability of any data relevant to their research questions,
- the availability of documentation on the relevant data and research projects,
- the complementarity and quality of the relevant data.

In addition, when planning a new survey:

- the availability of research instruments and information on specific methodological procedures used in similar research projects.

Keep in mind that if we are able to find useful external data and materials and incorporate them in our research study, then the proposal should also foresee the expected financial, temporal and organisational expenses of acquiring them.

2.2. ETHICAL AND LEGAL CONDITIONS OF WORK WITH SOCIAL DATA

There are numerous codes of ethics and sets of standards that apply to empirical social research. Among the most important ones are the ICC/ESOMAR[9] International Code on Market and Social Research [ESOMAR 2008] and the WAPOR[10] Code of Ethics [undated]. Additional codes of ethics are defined for the different disciplines of research based on social science data and many specific research areas. One is also incorporated in the international standard for market, opinion and social research, ISO 20252:2006. Such regulations cover relations between researchers and respondents, researchers and their clients or sponsors, and mutual relationships among researchers themselves. Since human society is the subject of research, rule-breaking harms not only the research study's direct participants, sponsors or clients, but also people's general willingness to participate in research. Thus, such malpractice may complicate the implementation of future research studies and ultimately undermine the credibility of the field as a whole.

The fact that research ethics are constructed in this way has numerous implications for our work with data. These are discussed comprehensively in the publications of the UK Data Archive [UKDA 2010], the European network CESSDA [2009] and Marcia Freed-Taylor [1994], among others. The main ethical requirements of data management, beyond the general requirements of the quality of scientific work, can be summarised as follows:

- Respondents should be protected from the potential harmful effects of research even after the stage of field data collection has been concluded, in particular whenever the data are worked with, archived, made available, or made subject to secondary analysis. In general, information of an individual nature about survey participants and other personal data is confidential, and this confidentiality should be maintained. Special attention should be paid to sensitive information.

[9] International Chamber of Commerce (ICC)/World Organization for Opinion and Marketing Research Professionals (ESOMAR).

[10] World Association for Public Opinion Research (WAPOR).

- Respondents must be treated with respect and have the right to know the purpose and methods of utilisation of the information they provide and to decide about the ways it can be utilised. Consequently, their decisions must be respected.
- Adequate utilisation of the information gathered in line with the purpose defined should always be ensured, not only to fructify the efforts respondents made to participate in the research study. Data gathered with public funding must be utilised as much as possible and, whenever the nature of the data allows, made available to the broader scientific community.

For some types of data, those rules are also provided by the laws. In particular the legal regulation of personal data protection represents an important factor shaping the possibilities of research work. It prescribes the obligation to obtain respondents' informed consent to the processing of their personal data and to maintain its confidentiality, and stipulates the methods for preventing the disclosure of personal data. Other legal norms that substantially shape data management include copyright and intellectual property laws. Such laws also appertain to data sharing. In some situations and for some specific exercises, the provisions of the Act on Archiving or other laws may be applicable as well.

2.3. PERSONAL DATA PROTECTION

The issues of personal data protection should be given adequate attention as early as the stage in which a research proposal is drafted. To underestimate them would not only constitute a violation of research ethics but might also restrict or completely prevent the researchers' intentions from being fulfilled and in particular the data from being made available for secondary research. The following should be clear from the beginning:

- Is it necessary to obtain respondents' informed consent for personal data operations?
- Will the data have to be anonymised?

A simple yes/no answer to these questions is insufficient; additional details are important. We need to exactly identify the phases of the research and the data life cycle in which the presence of personal information on respondents is unavoidable. Then we should plan our data management in such a way that it avoids any unnecessary operations with personal data and institutes adequate personal data protection measures where such operations cannot be avoided. Only such an approach allows us to meet the requirements of both research ethics and the law. Moreover, it may also simplify the survey process and promote research quality and efficiency.

In the following overview, I will draw on Czech legislation. On the one hand, legal regulation in European countries is to some extent similar because it is based on a common directive of the European

Union.[11] On the other hand, there are significant differences. For example, the Czech Republic does not apply specific rules to personal data processing for scientific purposes and its laws in this respect are among the strictest. The definition of personal data in Czech legislation is provided in Box 1. In social research, it must be carefully considered that data are not anonymous even if the data subject is identifiable only indirectly (see Box 2).

A general rule for any research study based on the collection of information from respondents is to obtain their informed consent. When the study deals with personal data this is also regulated by law. The data subject must at least be informed properly and in advance about the purpose of the data processing, the scope of the personal data, the name of the processor, and the time period the consent is given for. When it comes to so-called sensitive data, in practice consent must be obtained in writing and must preferably also signed by the respondent to demonstrate the existence of consent as required (see Boxes 1 and 3). The respondent is also entitled to request further information about the data processing, and if the reasons for which consent was obtained cease to exist, the data processor must stop processing, i.e. must liquidate the data.

In order to process sensitive data, the data processor must also register their activity with the Office for Personal Data Protection.[12] In other words, any institution planning to implement a research project that includes the processing of sensitive data must have a relevant reason for this kind of activity, have an adequate structure for securing the protection of such data, and must officially register in time, i.e. before the data processing begins. The implementation of this kind of data management itself entails, of course, additional organisational requirements and expenses.

For these reasons, it is necessary to consider carefully which exercises essentially depend on processing personal data and obtain informed consent and implement data protection measures for these exercises. However, even if social research tends to rely on the collection of individual data, it seeks to obtain aggregated information about society. Thus, personal data can often be omitted altogether or at least in some research stages. If this is the case, the data should be collected as anonymous or should be anonymised as soon as circumstances allow. Furthermore, there are also organisational reasons for doing so; informed consent is easier to obtain for a limited time period and a clearly defined purpose than for an extensive research exercise where respondents do not clearly understand the purpose or the consequences for them personally.

[11] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. In 2012 the European Commission proposed a new EU General Data Protection Regulation that should supersede the Data Protection Directive. The rules of personal data protection in research are considerably more stringent in this proposal, but in some aspects also closer to current Czech law.

[12] Office for Personal Data Protection in the Czech Republic: http://www.uoou.cz (visited 12 September 2013).

BOX 1: PERSONAL, SENSITIVE AND ANONYMOUS DATA

"personal data" shall mean any information relating to an identified or identifiable data subject. A data subject shall be considered identified or identifiable if it is possible to identify the data subject directly or indirectly in particular on the basis of a number, code or one or more factors specific to his/her physical, physiological, psychical, economic, cultural or social identity;'

"sensitive data" shall mean personal data revealing nationality, racial or ethnic origin, political attitudes, trade-union membership, religious and philosophical beliefs, conviction of a criminal act, health status and sexual life of the data subject and genetic data of the data subject; sensitive data shall also mean a biometric data permitting direct identification or authentication of the data subject;'

"anonymous data" shall mean such data that cannot be linked to an identified or identifiable data subject in their original form or following processing thereof;'

Source: Czech Republic [2011].

Czech Republic, Act No. 101/2000 Coll., Article 4

BOX 2: DIRECT AND INDIRECT IDENTIFIERS

Direct identifiers include names, national identification numbers, addresses etc. Indirect identifiers make a person's identification possible when associated with other known data. Such identification may be possible also by combining data from multiple different variables in the data file. For example, if the dataset contains information about a person's job as mayor in a given city, then a specific person holding that office can be identified even though the database does not contain that person's name.

BOX 3: PROCESSING OF SENSITIVE DATA

The following rules apply to the processing of sensitive data in research. Sensitive data may be processed:

'if the data subject has given his express consent to the processing. When giving his consent, the data subject must be provided with the information about what purpose of processing, what personal data, which controller and what period of time the consent is being given for. The controller must be able to prove the existence of the consent of data subject to personal data processing during the whole period of processing. The controller is obliged to instruct in advance the data subject of his rights pursuant to Articles 12 and 21…'

Source: Czech Republic [2011].

Czech Republic, Act No. 101/2000 Coll., Article 9

Note: Article 12 refers to data subjects' access to information, Article 21 to the protection of their rights.

I will illustrate this with a normal quantitative survey of the general population. While random sampling tends to identify specific addresses, the research study itself can make do without direct identification of households and respondents. Therefore, the dataset does not have to include such direct identifiers. If the database does not include so-called indirect identifiers (see Box 2) either, then it is anonymous and no informed consent is required for analysing, archiving or sharing the data. Similarly, in a panel survey, we need to preserve the addresses in order to implement follow-up waves and survey the same units, but not for the analysis itself. Thus, addresses can be kept separately from the data collected. The database of addresses will be treated in line with the Personal Data Protection Act, while the dataset for analysis will remain anonymous.

Another situation frequently arises: at a certain stage, often just for the purposes of data collection or building connections between databases, personal data has to be used, but in all the other stages of the research the personal data of respondents can be omitted – and discarded. The process of discarding identifiers is referred to as anonymisation and its methods are described below. At this point, we should add that even in the above-described case respondents' informed consent must be obtained for those exercises where we work with non-anonymous data.

If the database contains direct or indirect identifiers and cannot be anonymised, we must obtain informed consent from the respondents and count on spending money on personal data protection. If the database is not anonymous, given the topics typical for social research, which often involve sensitive data, in the Czech Republic consent should be given in writing and data protection measures must be more thorough. At the same time, this poses an important barrier to data sharing. The purpose of data processing must be formulated in specific and time-limited terms in the informed consent request. In the Czech Republic, it is impossible to obtain consent to the processing of data for an unlimited time, for an unknown purpose, or for the purpose of sharing it with anybody. As a result, non-anonymous databases are usually not made available in data archives.[13]

[13] However, institutions holding the respondents' consents with personal data processing may share such data with other researchers via '"remote access"'. The researchers will be provided with the results of analyses but will not come in touch with primary data. It should be also mentioned that some institutions, in the Czech Republic namely the Czech Statistical Office and the Institute for the Study of Totalitarian Regimes, are governed by specific laws including personal data processing stipulations.

2.4. COPYRIGHT AND INTELLECTUAL PROPERTY PROTECTION

Social scientists normally deal with copyright and intellectual property rights (IPR) when it comes to the publication process or the publicising or commercialising of some other types of research results. However, people pay too little attention to and have ample misconceptions about copyright issues when it comes to databases and data sharing. Data processing tends to be governed by customs that differ from discipline to discipline. Moreover, some of these customs contradict the legal order or the ownership of copyright to some databases is unclear, their users have no knowledge of copyright and they ignore it when conducting their research. In this respect, some institutions set up dysfunctional types of internal copyright regulations pertaining to databases. Then, if they are obliged to abide by them, for example, in the case of a dispute, they can easily find themselves unable to fulfil their commitments. In some cases faulty legal treatment may even result in the researchers' total failure to fulfil their project's goals, and typically it poses unnecessary barriers to making data available for secondary use.

Copyright and intellectual property protection is complicated and a thorough treatment of these issues requires professional legal advice. In each research institution such legal advice should result in the creation of ground rules and standard practices for its employees to follow. Nevertheless, each researcher should be aware of at least the basic contexts. I shall provide some useful information to this end. With regard to differences between the environments of law and research, some data organisations have published copyright guides that are relevant to their respective domains of interest. For example, UKDA published an online review of copyright issues intended for social science data services users [UKDA 2010: Copyright], and database administrators can also make good use of materials published by the JISC and the TLTP [JISC, TLTP 1998], British institutions supporting information system development.[14] In the Czech Republic, intellectual property rights are treated in particular in the Copyright Act, i.e. Act No. 121/2000 [Czech Republic 2006]. One must bear in mind the differences between national laws, but general principles are often similar.

Copyright covers any works which are the unique outcome of the creative activity of the author and are expressed in any objectively perceivable manner including electronic form, permanent or temporary, irrespective of their scope, purpose or significance (see Box 4). A database which, by way of the selection or arrangement of its content, is the author's own intellectual creation, and in which the individual parts are arranged in a systematic or methodical way and are accessible by electronic or other means, is a collection of works, and as such it is covered by the Copyright Act. Copyright arises when the database is being created. The fact that a database does not bear a 'copyright' label does not exclude it from this legal framework.

[14] Joint Information Systems Committee (JISC), Teaching and Learning Technology Programme (TLTP).

BOX 4: AUTHOR'S WORK

'The subject matter of copyright shall be a literary work or any other work of art or a scientific work, which is a unique outcome of the creative activity of the author and is expressed in any objectively perceivable manner including electronic form, permanent or temporary, irrespective of its scope, purpose or significance...'

Czech Republic, Act 2000/121 Coll., Article 2 (1)

'...A database which by the way of the selection or arrangement of its content is the author's own intellectual creation, and in which the individual parts are arranged in a systematic or methodical way and are individually accessible by electronic or other means, is a collection of works...'

Source: Czech Republic [2006].

Czech Republic, Act 2000/121 Coll., Article 2 (2)

Copyright protection covers the authors' work, not the individual facts stated in it. As far as databases are concerned, this means that copyright covers the selection and arrangement of data in a database etc., while its content may not be covered, depending on what exactly the content is. For example, for an in-depth interview, copyright to the recording is held by the researcher, while the rights to the individual statements remain with the informant. Therefore, besides personal data protection issues (see point 2.3 above) this is another reason for a researcher who is creating a database of qualitative data to obtain the informant's written consent.

Copyright protects intellectual property from unauthorised distribution, given the potential loss of income and moral damage. The rightholder chooses the ways of disposal of his/her work and decides about its distribution. Nevertheless, copyright is not infringed by anybody who in his or her own work to a justified extent uses excerpts from the work of other authors, or small works in their entirety, for the purposes of the critique or review of such a work or for the purposes of scientific or technical work, or uses the work while teaching for illustrative purposes or in non-commercial scientific research. However, in doing so, it is always necessary to cite the name of the author, the title of the work, and the source.

Through a written license agreement, the author can grant authorisation to use the work, either in specific ways or in all ways of use, and either to a limited or to an unlimited extent. A license can be either exclusive or non-exclusive. In the case of an exclusive license, the author must refrain from further distribution of and from exercising the rights to use the work to which he granted the license.

The copyright belongs to all the authors of the work, for example, the entire research team, and not only the team leader or the project's principal investigator. The same applies to university research: the rights do not belong to the teacher only but also to the students who participated in organising the

research study. However, a person who has contributed to the creation of the work merely by providing assistance or advice of a technical, administrative or expert nature or by providing documentation or technical material, or who merely gave the impulse to create the work is not considered to be a joint author.

Databases are often created in the framework of an employment relationship. As a rule, the employer exercises the author's economic rights to a work in his or her own name. Economic rights cover the different ways of using the work, e.g. reproduction, distribution, exhibition, lending, or making the work available. The author's moral rights, e.g. the right to claim authorship, the right to the inviolability of a work (alterations), or the right of supervision over compliance with obligations, remain unaffected.

Thus, authorisations for the secondary use of or access to a database in an archive are often granted by the employer, rather than the authors' team. In this respect, it is worth mentioning that most students are not employees of their university, which means that economic rights to their works are not transferred to the university in their entirety.[15] In some cases, too, academic institutions transfer economic rights to their employees, especially for the purposes of publication activity; sometimes the scope of these institutional rules includes other outcomes and activities as well, which may affect the regulation of rights to databases.

Databases can also be created and shared in an environment of wide-open collaboration based on free licenses such as Creative Commons. Then, users can not only utilise the database but can also contribute to it, expand it, update it, or make other alterations, subject to license conditions.

[15] The student-author-university relationship is more complex. Schools or school-related or educational establishments have the right to conclude, under the usual terms, a license agreement on the utilisation of a school work. Unless otherwise agreed, the author of a school work may use his work or may grant the license to any other party, unless this contravenes the legitimate interests of the school.

2.5. DATA MANAGEMENT PLANNING

Data management takes place in several phases of a research study and includes numerous, often interrelated, elements and processes. Omission of these elements and processes might result in significant negative effects for the utility of a database and for the course and results of the research process. For these reasons, we should proceed in a systematic and planned manner. In order to clarify the basis of a research project before we begin collecting data, it is recommended that researchers draft a 'data management plan' that answers some basic questions affecting the course of the research process. This document summarises how the data were generated and handled both during the research process and after its conclusion.

A data management plan may be highly formalized, which depends on another potential function it has. Many funding agencies require researchers to draft such a plan in their grant application and use it to assess and check the project's compliance with the principles of open access to data. This is the practice of the US National Science Foundation (NSF), UK Research Councils and many other funding agencies.[16]

The drafting of the plan, whether a funding agency wants it or not, can be done according to one of the following models, which provide a checklist of items that should not be omitted from the plan. Since the elements of a data management plan may vary depending on research objectives and/or funding agency requirements, we should select and use the models with discretion and reflect the specific research situation in our planning.

The ICPSR archive [ICPSR 2011] provides extensive online services for data management planning, with special emphasis on the social sciences.[17] Its website contains a detailed set of instructions, including a checklist of elements of a data management plan and a recommended framework for drafting it, many examples and models of data management plans, and other instructions. The Data Curation Centre[18] provides advice for data management planning in a multidisciplinary context, with close attention paid to the requirements of funding agencies in the UK. Table 2 presents a schematic and modified version of the ICPSR framework for preparing data management plans.

[16] See, for example, NSF Data Archiving Policy: http://www.nsf.gov/sbe/ses/common/archive.jsp; a summary of the UK Research Councils' requirements: *http://www.dcc.ac.uk/webfm_send/358*; or the ESRC guidelines for peer reviewers of data management plans: *http://www.esrc.ac.uk/_images/Data-Management-Plan-Guidance-for-peer-reviewers_tcm8-15569.pdf* (all sites visited 12 September 2013).

[17] See ICPSR [2011] for ICPSR Guidelines for Effective Data Management Plans in the form of a document; see *http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/index.html* for an online presentation (visited 12 September 2013).

[18] Data Management and Sharing Plans (DMPs) at the Data Curation Centre (DCC): *http://www.dcc.ac.uk/resources/data-management-plans* (visited 12 September 2013).

TABLE 2: ELEMENTS OF A DATA MANAGEMENT PLAN RECOMMENDED BY THE ICPSR

Element	Description
Data description	A description of the information to be gathered; the nature and scale of the data that will be generated or collected.
Existing data	A survey of existing data relevant to the project and a discussion of whether and how these data will be integrated.
Format	The formats in which the data will be generated, maintained, and made available, including a justification for the procedural and archival appropriateness of those formats.
Metadata	A description of the metadata to be provided along with the generated data, and a discussion of the metadata standards used.
Storage and backup	Storage methods and backup procedures for the data, including the physical and cyber resources and facilities that will be used for the effective preservation and storage of the research data.
Security	A description of technical and procedural protections for information, including confidential information, and how permissions, restrictions, and embargoes will be enforced.
Responsibility	The names of the individuals responsible for data management in the research project.
Intellectual property rights	Entities or persons who will hold the intellectual property rights to the data, and how IP will be protected if necessary. Any copyright constraints (e.g. copyrighted data collection instruments) should be noted.
Access and sharing	A description of how data will be shared, including access procedures, embargo periods, technical mechanisms for dissemination and whether access will be open or granted only to specific user groups. A timeframe for data sharing and publishing should also be provided.
Audience	The potential secondary users of the data.
Selection and retention periods	A description of how data will be selected for archiving, how long the data will be held, and plans for the eventual transition or termination of the data collection in the future.
Archiving and preservation	The procedures in place or envisioned for the long-term archiving and preservation of the data, including alternative plans for the data should the expected archiving entity go out of existence.
Ethics and privacy	A discussion of how informed consent will be handled and how privacy will be protected, including any exceptional arrangements that might be needed to protect participant confidentiality, and other ethical issues that may arise.
Budget	The costs of preparing data and documentation for archiving and how these costs will be paid. Requests for funding may be included.
Data organisation	How the data will be managed during the project, with information about version control, naming conventions, etc.
Quality Assurance	Procedures for ensuring data quality during the project.
Legal Requirements	A listing of all relevant federal or funder requirements for data management and data sharing.

Source: ICPSR [2011].

2.6. BUDGETING

Data management entails financial expense. It is advisable not to ignore this fact in project budgeting. The UKDA has prepared a special instrument for planning data management expenses [UKDA 2011]. It was designed for the British Economic and Social Research Council (ESRC) and social scientists applying for funding from its programmes. In it, budgeting includes the following activities that may be relevant for different projects:

- obtaining informed consent
- anonymisation
- data security and access (unauthorised access, personal data protection)
- digitisation
- transcription (e.g. of interviews)
- formatting and organising files (formatting and changes in the arrangement of databases)
- data labelling and coding
- cleaning
- data context description
- documentation (obtaining documentation during or after the process)
- metadata (creating data description/documentation)
- file format (costly conversion of audiovisual data etc.)
- planning, distribution of roles and responsibilities (collaboration between multiple institutions etc.)
- operationalising (data management planning and implementation)

3.1. DATA FILE STRUCTURE

Data files may have different internal structures and a research study may encompass several data files in different relations to one another. The structure of a data file is also determined by the formatting and arrangement of variables. File structure choice largely depends on the requirements of the software we are using. On the other hand, decisions about structure that are necessarily taken when a data file is being created are going to define the possibilities of data processing and data analysis; once the structure has been filled with data, any changes to it are usually laborious and costly.

Therefore, the initial decisions we make about data structure should be considered thoroughly. We should clarify the unit of analysis and define database records in line with the structure. We should take into consideration our analytical objectives, as well as the methods and software we are going to apply. Special attention should be paid to any specific analytical procedures and/or specific software applications for implementing them. Additionally, our decisions will depend on the relations between items as well as on our assumptions about the variables that will be created during data entry and about the relations between them. Another factor is our assumptions about follow-up surveys and additional data sources, their structure, and the possibilities of interconnection. We may add new data to our collection, create cumulative files, or build interconnections with other databases and information sources. We will probably create different versions of the file in the course of our own analysis as well.

Good reasons for choosing a certain structure may also be related to the number of variables, the number of cases, or the total size of the database. Some software applications have strict limits in this respect; clarity and operability during processing may pose another problem. Even when using very recent technologies, we may be faced with limits to available computing or storage capacity. Another issue is the blank cells in the file, which are multiplied geometrically in some file structures, resulting in the excessive growth of the size of the file.[19] The following basic types of data files are distinguished in terms of structure [see, e.g., ICPSR 2009: 26–27]:

- **Flat file:** the data are organised in long rows, variable by variable. An ID number usually comes first. If variable values are organised column by column, we obtain a rectangular matrix. This is the simplest structure. For example, SPSS system files consist of one rectangular matrix with data, accompanied by variable labels and values.

[19] Given the structure and variable format selected, we often reserve a large space in the file for information whose slots will remain blank. An example of a structure with large size requirements is a hierarchically structured database of households which foresees up to eight members for each household and defines variables and positions in the file for each of those members. Most households have less than eight members, while the variables and positions remain in the file even if they are blank.

- **Hierarchical file:** The file contains higher-order and lower-order records that are arranged in a hierarchical structure, i.e. several lower-order units may be linked to one higher-order unit. Such a structure may be used, for example, for household surveys where data on the household are recorded at one level and data on household members at another level. Different database applications such as MS Access or D-base often structure data in this way.
- **Relational database:** This is a system of data matrices and defined associations between them. For example, in a household survey, information about household members may be recorded in independent matrices that are interconnected by means of a household ID or a more complex parameter that represents not only the sharing of a household but also the type of family relationship between household members. For instance, users can search for rows with equal attributes in this type of database. Relational databases may also serve as a basis for creating files adapted for individual exercises by combining information from different matrices.

If we wanted to use a flat file for the above-mentioned household data exercise, we could add to the file a household ID variable and then organise records about the respondent and other members of his/her household row by row. This would create a set of individuals. Another possibility would be to organise records for all household members in long rows, which would create a set of households. Another alternative would be to create several interconnectable files.

The requirements and abilities of our software often dictate whether we create one compact but complicated and sizable file, or several simpler and smaller interconnectable files. In addition, large surveys often use several different databases. These may consist, for instance, of files from different waves of a survey, of databases with different types of research units, or, to give a concrete example, of data from the main questionnaire, different supplements, contact forms, contextual data etc. As a result, decisions about them are more complicated. A specific situation exists for data obtained from continuous surveys or several different sources. We should always define reliable and unique identifiers of records in different files so that we can interconnect files where links exist between them.

3.2. VARIABLES

The file's structure is further shaped by the arrangement and labelling of variables. There are relationships between variables, e.g., an original and a derived variable or sets of variables linked to the measurement of the same phenomenon. At the same time, there are links to other elements of the research study such as the questionnaire, data source, or another dataset. Besides variables which comprise directly measured data, there are also derived variables. This all should be reflected in the location of variables in the data file and their labelling in order to help users better understand the contents of a database, gain orientation in it during analysis and avoid some mistakes.

Most data files also include auxiliary variables which facilitate orientation and management, ensure integrity, or are necessary for some analyses. As a rule, we should include a unique identifier of cases in the file and place it in the very beginning of it. Other variables may help us distinguish between different sources of information, methods of observation, temporal or other links. Yet others may provide information about the organisation of data collection such as interviewer ID or interviewing date, or distinguish cases which belong to various groups. It is absolutely necessary for an analysis to distinguish data that result from overrepresentation sampling strategies, different surveying waves etc., especially if groups of cases distinguished by them are to be analysed in different ways. See Table 3 for an example of the arrangement of variables in the data file.

TABLE 3: STRUCTURE OF THE CZECH NATIONAL DATA FILE FROM THE SURVEY ON SOCIAL INEQUALITY 2009 CONDUCTED WITHIN THE FRAMEWORK OF THE INTERNATIONAL SOCIAL SURVEY PROGRAMME (ISSP)

Variables	Description
IDRESPO	Identification of the respondent
DAY, MONTH, TIME_ST, TIME_END	Variables identifying the date and timing of the interview
Q1a to Q20c	Variables of the ISSP module section of the questionnaire
P1 to P4	Variables of the national specific section of the questionnaire on political attitudes
S1 to S39	Variables of the national specific socio-demographic section of the questionnaire
T1 to T3	Items of the questionnaire answered by the interviewer (sex, district of residence, region)
W1	Design and post-stratification weight
AGE, AGECAT, ISCO88 to CLS6_MO	Derived variables (categorised and standardised versions of original variables)

Source: Institute of Sociology ASCR, Czech Social Science Data Archive.

The naming and labelling of variables is as important as their location in the data file. Since variable names are also used as calling codes in software operations, they should be kept short and respect the usual requirement of standard software. Although the limits of contemporary software are usually looser, abidance by standards and thorough avoidance of specific arrangements are necessary for the reasons of transferability, long-term data preservation, and the prevention of unnecessary costs due to format conversions. Therefore, variable names should be no longer than eight characters, should start with a letter, not a number or other characters such as question marks or exclamation marks, and should not contain special characters such as #, &, $, @, which are often reserved for specific purposes in software applications. Do not use diacritics or national specific characters under any circumstances.

At the same time, variable names should not be completely meaningless since they can be used for better orientation in the file. Three basic ways of variable labelling are customary:

- a numeric code that reflects the variable's position in a system (e.g. V001, V002, V003...),
- a code that refers to the research instrument (e.g. question number in a questionnaire: Q1a, Q1b, Q2, Q3a...),
- mnemonic names referring to the content of variables (BIRTH for year of birth, AGE for respondent's age etc.), sometimes with prefixes, roots and suffixes to distinguish variables' membership in groups or links between them (e.g., AGECAT for a categorised variable derived from AGE).

Specialised statistical software normally makes it possible to use not only variable names but also variable labels, which can be longer and help us better specify the variable's content. Here we can enter more complete information, e.g. a short or full version of the question, or alternative labels such as question codes if the variable names are not constructed around them. Although more emphasis is placed on comprehensibility, and size limits are less strict here, it is advisable to keep variable labels rather brief or to find an adequate compromise. Excessively lengthy labels make analytic outcomes unclear and complicate format conversion. In some uses or after conversion between software applications, only a part of a lengthy label is kept and the loss of the rest of it may be to the detriment of comprehensibility. The use of diacritics should be considered carefully because they tend to cause similar problems. On the one hand, they increase the quality and comprehensibility of analytic outcomes written in Czech, but on the other hand, they may complicate conversion between software formats or work with results in a different language version of the same software.

Below I will discuss two examples of variable labelling for survey data. Tables 4 and 5 outline variable names in datasets from the International Social Survey Programme (ISSP). The first case describes the Czech dataset from ISSP 2010 entitled, 'Environment III'. The list begins with an ID variable. Variable names correspond to the codes of the questions in the questionnaire (easy orientation, direct reference to the questionnaire), while variable labels copy the wording of questions, excluding diacritics (easier transferability, almost complete wording of questions available, but the labels are quite lengthy).

TABLE 4: EXAMPLE OF VARIABLE LABELLING: EXCERPT FROM THE VARIABLE LIST OF THE CZECH ISSP 2010 DATASET ON 'ENVIRONMENT III' (TRANSLATED INTO ENGLISH)

Variable name	Variable label
id	Questionnaire No.
q1a	Which of these issues is the most important for the Czech Republic today?
q1b	Which is the next most important?
q2a	Private enterprise is the best way to solve the Czech Republic's economic problems
q2b	It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes
q3a	The Czech Republic's highest priority should be to…
q3b	And what is the Czech Republic's next highest priority, the second most important thing it should do?
q4	Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?
q5a	Most of the time we can trust people in government to do what is right
q5b	Most politicians are in politics only to get something out of it for themselves
q6	How concerned are you about environmental issues?
q7a	What problem do you think is the most important for the Czech Republic as a whole?
q7b	What problem affects you and your family the most?

Source: Institute of Sociology ASCR, Czech Social Science Data Archive.

The second case (Table 5) describes the international dataset from ISSP 2009 on 'Social Inequalities'. It combines two approaches to variable labelling. The first, thematic part of the file contains simple variable names (numeric codes). However, variable labels begin with the numbers of the questions in the questionnaire – because there is an international master questionnaire and the numbers may not correspond to national versions of the questionnaire. What follows is an incomplete English version of the question, which was shortened adequately to remain comprehensible and keep the variable label short.

Some ISSP surveys allow alternative wording of questions – possible alternatives are bracketed in inequality signs. Similarly, after country specifics (e.g., country name, currency used), general names come in inequality signs. The second part of the ISSP file, which contains sociodemographic variables that are not directly linked to the wording of questions in the international questionnaire but instead constructed from national versions of data, uses mnemonic names of variables referring to their contents and simultaneously to links between them (e.g., DEGREE = the education variable transformed into an internationally comparable form, XX_DEGR = education variables using original country-specific coding).

TABLE 5: EXAMPLE OF VARIABLE LABELLING: EXCERPT FROM THE VARIABLE LIST OF THE INTERNATIONAL ISSP 2009 DATASET, 'SOCIAL INEQUALITIES'

Variable name	Variable label
V73	Q24a Describe yourself: I work hard to complete my daily tasks
V74	Q24b Describe yourself: I perform to the best of my ability
V75	Q24c Describe yourself: I work hard to maintain my performance on a task
V76	Q25a Describe yourself as <14-15-16> years old: I tried hard to go to school everyday
V77	Q25b Describe yourself as <14-15-16> years old: I performed to the best of my ability
SEX	R: Sex
AGE	R: Age
MARITAL	R: Marital status
COHAB	R: Steady life-partner
EDUCYRS	R: Education I: years of schooling
DEGREE	R: Education II: highest education level
AR_DEGR	Country specific education: Argentina
AT_DEGR	Country specific education: Austria
AU_DEGR	Country specific education: Australia
BE_DEGR	Country specific education: Belgium

Source: International Social Survey Programme (ISSP), GESIS Data Archive.

Another exercise consists of setting the format of variables. First, we have to determine an adequate type of variable, choosing between string and numeric variables and, in some software applications, between nominal, ordinal, ratio or interval variables. Second, we need to set the variable length, i.e. the number of characters or the length of the integer and fractional parts of a number. The setting must correspond to the nature of the data and affects the possibilities of analysing data and displaying results. Variable format also affects the file size, since the set number of characters or digits for each variable is reserved for every case, even if they are left blank. Thus, given hardware limitations, inadequate variable formats may slow down processes or put excessive demands on computing or storage capacity.

3.3. VARIABLE VALUES, CODING

There are different types of variable values. Variables may refer to numeric values, but most variables in a sample survey contain respondents' oral responses to open-ended or closed-ended questions in the questionnaire. Additionally, datasets may contain photographs, video recordings, audio recordings or samples of different materials.

For the purpose of quantitative analysis, the information collected is usually represented by numeric codes. The fact of numeric coding is shared by all statistical software applications and, among other things, this facilitates data conversion and measurement comparisons. Our decisions about the specific structure of coded categories also affect the prevention of errors in data entry and data processing and define the basis of analysis. Generally speaking, coded categories should refer to the contents of the hypotheses tested [e.g., Groves et al. 2004: 306–307]. At the same time, one must take into consideration any possible uses of the data in future and maximise the data's informational value.

The meaning of codes must be documented. Specialised analytic software lets the user assign value labels directly to variable values. The construction of value labels follows principles similar to those of variable labels. We should make them comprehensible but avoid wasting room in order to maintain the clarity of the analytic outcomes using them and the transferability of information between software. If the application does not allow us to assign codes directly to data, we have to document the values in a separate file or as part of the database's more general documentation file.

In practice, coding takes several different forms. For closed-ended questions, the coding scheme is incorporated directly in the questionnaire and data are entered numerically. This process is automated in computer-assisted interviewing. More complex coding exercises, e.g. for textual answers to open-ended questions, require an independent coding process with a clearly defined design: a coding structure and a procedure and schedule of exercises if there are several coders. Answers can also be coded on paper questionnaires, when coders record codes in a designed spot of the questionnaire before they are entered into the computer. Such codes are then digitised along with other data. It is more advisable to enter the complete wording of textual answers and then conduct coding in the computer. This enhances control both during and after the process, facilitating the correction of mistakes or the making of subsequent changes to the coding system. Sometimes, a part of the process can be automated as well (e.g. recommended procedures for coding using the ISCO classification [Ganzeboom 2010]).

For example, the following coding recommendations are included in the standard for creating data files, which is part of the ICPSR's [2009: 15–16] manual for data depositors. I have slightly modified and simplified them for the purposes of this paper:

- Identification variables: Include identification variables at the beginning of each record to ensure unique identification of each case.
- Code categories: Code categories should be mutually exclusive, exhaustive, and precisely defined. You should be able to assign each response of the respondent into one and only one category.
- Preserving original information: Code as much detail as possible. Subsequently, your information can be converted to a less-detailed one, but this cannot be done the other way around. Less detail will limit the possibilities of secondary analysts.
- Closed-ended questions: For responses to survey questions that are coded in the questionnaire, you should digitise the coding scheme to avoid errors.
- Open-ended questions: Any coding scheme applied should be reported in the documentation.
- Recording responses as full verbatim text: Such responses must be reviewed for personal data protection.
- Check-coding: It is advisable to verify the coding of selected cases by repeating the process with an independent coder. This checks both the coder's work and the coding scheme.
- Series of responses: If a series of responses requires more than one field, it is advisable to apply a common coding scheme distinguishing between major and secondary categories etc. The first digit of the code identifies a major category, the second digit a secondary category etc.

Equal coding structures can be used for several variables in one research study. I would like to add to the above list of recommendations that coding schemes for each variable should be constructed in consideration of the rest of the database, so that the character of the procedures does not deviate from each other excessively. For example, the response scales constructed should have the same direction. Such a rule facilitates coding, data processing and the prevention of errors (both in data entry and in analysis). Moreover, it motivates us to formulate our coding practices as early as in the stage of designing the questionnaire (see above).

The coding of answers to fully standardised questions is usually a simple exercise. Yet open-ended questions may pose a methodological challenge. We must decide to what extent different answers to the same question are equivalent and can be assigned to the same categories. Thus, coding is a complicated cognitive process and the coder may exert a significant influence on the information that appears in the database, as well as a source of substantial mistakes and systematic errors of measurement. Moreover, in some cases, we need to code information in great detail, using a complicated scheme. Thus, the design of coding schemes requires a theoretically and empirically well-founded planning and testing. Similarly, coding procedures must be planned, which requires the establishment and implementation

of a specific coding design and places specific demands on coders' competences and training. A part of the coding procedures is concerned with reviewing the quality of the coding process, including, for instance, the assessment of coder variance.[20]

The demands of research in this respect are somewhat alleviated by the fact that the same coding structures can be applied in different studies. The use of standardised classifications and coding schemes has multiple advantages, in particular (1) economy and quality as a result of adopting an existing structure which has a solid basis and has been verified in many studies, (2) comparability with data from other studies using the same concept, (3) comprehensibility for researchers used to working with these concepts. A disadvantage lies in the need to adapt one's research intent.

Occupational classifications such as the International Standard Classification of Occupations (ISCO) are examples of complex coding systems and widespread standard coding schemes. Occupational information has several dimensions and needs to be collected in relatively extensive detail. This is, as a rule, done by means of one or more open-ended questions. The above-mentioned national and international classifications are based on similar theoretical and methodological concepts and are highly extensive. They use four- and five-digit codes, within which they are structured and generic – covering several levels of generalization. There are specific rules and recommendations for implementing the coding process, for instance that the information should be gathered in string and numeric forms simultaneously, multiple coders should be employed, and their functions should overlap to enhance control [e.g. Ganzeboom 2010].

Standardised classifications are shared by many subjects. A large number of standardised classifications for different purposes that are adapted for Czech society and can be used for coding are available on the Czech Statistical Office's website in the section 'Classifications, Nomenclatures'.[21] When I was writing this text, a specialised database of metainformation was being established in which these classifications and nomenclatures were to be ordered systematically. Internationally applicable versions of classifications and nomenclatures are available on the websites of international organisations.[22] Furthermore, certain specialised projects and activities aim at standardising and harmonising international data in various areas. Examples include Harry Ganzeboom's website on stratification research[23] or the Portal of European Social Indicators managed by the German GESIS.[24] Box 5 contains an excerpt from the ISCO-2008 classification of occupations.

[20] Coder variance is a part of the total variance of values measured which can be attributed to individual coders' different ways of using the coding scheme [Groves et al. 2004: 316].

[21] CZSO website: *http://www.czso.cz* (visited 12 September 2013).

[22] For example, the ISCO at the website of the International Labour Organization (ILO), *http://www.ilo.org/public/english/bureau/stat/isco/index.htm,* or the ISCED-97 classification of education at the UNESCO website, *http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx* (visited 12 September 2013).

[23] *http://home.fsw.vu.nl/hbg.ganzeboom/* (visited 12 September 2013).

[24] *http://www.gesis.org/en/services/data-analysis/social-indicators/portal-european-social-indicators/* (visited 12 September 2013).

BOX 5: EXAMPLES OF CODES FROM THE ISCO-2008 CLASSIFICATION OF OCCUPATIONS

2	Professionals
21	Science and engineering professionals
211	Physical and earth science professionals
2111	Physicists and astronomers
2112	Meteorologists
2113	Chemists
2114	Geologists and geophysicists
212	Mathematicians, actuaries and statisticians
2120	Mathematicians, actuaries and statisticians
213	Life science professionals
2131	Biologists, botanists, zoologists and related professionals
2132	Farming, forestry and fisheries advisers
2133	Environmental protection professionals
214	Engineering professionals (excluding electrotechnology)
2141	Industrial and production engineers
2142	Civil engineers
2143	Environmental engineers
2144	Mechanical engineers
2145	Chemical engineers
2146	Mining engineers, metallurgists and related professionals
2149	Engineering professionals not elsewhere classified

Source: International Labour Organization.

3.4. MISSING VALUES

Not all the questions in a questionnaire are answered by all respondents, which results in missing values for the corresponding variables. It is advisable to distinguish between various missing data situations [ICPSR 2009: 17]. As a rule, the following situations are distinguished (frequently used acronyms are bracketed):

- No answer (NA): The respondent did not answer a question when he/she should have.
- Refusal: The respondent explicitly refused to answer.
- Don't Know (DK): The respondent did not answer a question because he/she had no opinion or did not know the information required for answering. As a result, the respondent chose 'don't know', 'no opinion' etc. as the answer.
- Processing Error: The respondent provided an answer but, for some reason (interviewer error, illegible record, incorrect coding etc.), it was not recorded in the database.
- Not Applicable/Inapplicable (NAP/INAP): A question did not apply to the respondent. For example, a question was skipped following a filter question (e.g. respondents without a partner did not answer partner-related questions) or some sets of questions were only asked of random subsamples.
- No Match: In this case, data are drawn from different sources, and information from one source cannot be matched with a corresponding value from another source.
- No Data Available: The question should have been asked, but the answer is missing for a reason other than those above or for an unknown reason.

Not all these missing data situations tend to be distinguished. It is crucial for the integrity of a database to at least document which questions were not asked of some respondents and specifically which respondents were not asked those questions. During coding, this information should be adequately recorded as 'Not Applicable/Inapplicable' (NAP, INAP). Furthermore, it is useful for many analyses to distinguish whether the respondent did not know the answer or simply did not answer (or refused to).

In order to facilitate data processing and error prevention, it is advisable to establish a uniform system for coding missing values for the entire database. Typically, negative values or values like 7, 8, 9, 97, 98, 99 (where the number of digits corresponds to the variable's format and number of valid values) are used for these purposes. The coding scheme should prevent overlapping codes for valid and missing values. For instance, whenever the digit zero is used for missing values, we should bear in mind that zero may represent a valid value for many variables such as personal income.

The examples in Boxes 6 and 7 refer to the preceding three sections on variables and their values. They are adopted from the Data Protocol of the fourth wave of the European Social Survey, which includes a detailed description of the survey's data files, their structure and contents. Box 6 contains an excerpt

from the variable list and complete information about the individual variables in the international database: the codes of the questions in the international questionnaire, the variable names (mnemonic names), the variable labels, the codes and value labels, including systematically coded missing values (6, 7, 8, 9 etc.), and notes outlining who did not answer a question and how the different categories of people are coded for that purpose. This information in the data protocol is accompanied by a diagram illustrating different questioning paths, which is presented in Box 7. In our case, question F13 (variable EMPLREL) was only answered by self-employed persons (value 2 in the EMPLREL question), and question F14 by all respondents other than self-employed.

BOX 6: EXCERPT FROM THE DATA PROTOCOL OF THE FOURTH WAVE OF THE EUROPEAN SOCIAL SURVEY: DOCUMENTATION OF VARIABLES

Table F.1f. Data file 1: Main questionnaire, section F

Qno	Name	Label	Format	Values	Categories	Comment
				9999	No answer	
F12	EMPLREL	EMPLOYMENT RELATION	F1.0	1	Employee	Ask F12 if F8a PDWRK=1 or F9=1 or F10=1 Go to F14
				2	Self-employed	Ask F13
				3	Working for own family business	Go to F14
				6	Not applicable	
				7	Refusal	Go to F14
				8	Don't know	
				9	No answer	
F13	EMPLNO	NUMBER OF EMPLOYEES RESPONDENT HAS/HAD	F5.0	66666	Not applicable	Ask F13 if F12=2. Go to F15 if number of employees given at F13.
				77777	Refusal	Go to F15
				88888	Don't know	
				99999	No answer	
F14	WRKCTRA	EMPLOYMENT CONTRACT UNLIMITED OR LIMITED DURATION	F1.0	1	Unlimited	Ask F14 if F12=1,3,7,8 Ask F15
				2	Limited	
				3	No contract	
				6	Not applicable	
				7	Refusal	Ask F15
				8	Don't know	
				9	No answer	

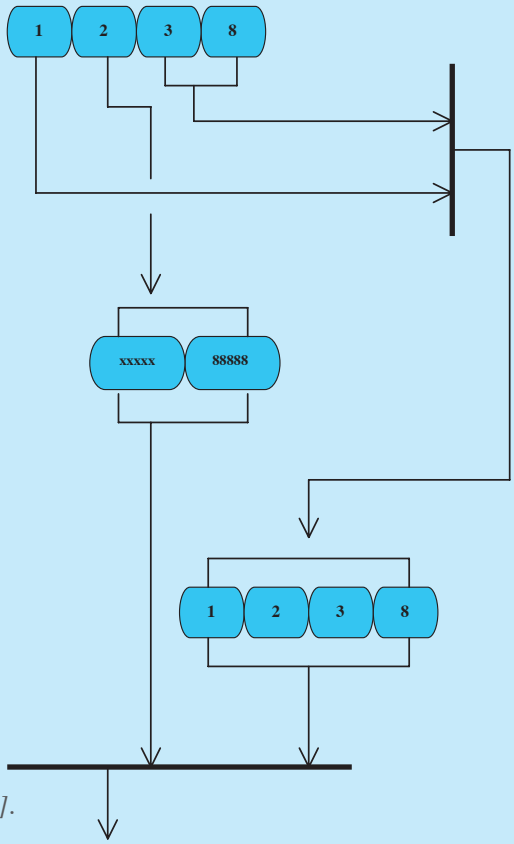
Source: European Social Survey [ESS 2008a: 45].

BOX 7: EXCERPT FROM THE DATA PROTOCOL OF THE FOURTH WAVE OF THE EUROPEAN SOCIAL SURVEY: SCHEME OF FILTRES AND THE QUESTIONING PROCEDURE

F12 EMPLREL
ASK F12: If code 01 at F8a or code 1 at F9 or code 1 at F10
GO TO F13: If code 2 at F12
GO TO F14: If code 1, 3,(7) or 8 at F12

F13 EMPLNO
ASK F13: If code 2 at F12
GO TO F15:
If number coded or code (77777) or 88888 at F13

F14 WRKCTRA
ASK F14: If code 1, 3,(7) or 8 at F12
GO TO F15: If code 1 to 3,(7) or 8 at F14



Source: European Social Survey [ESS 2008a: 76].

3.5. DATA ENTRY AND DATA FILE INTEGRITY

The goal is to enter data in a digital format that is fit for analysis and, in this process, to minimise errors such as typos and meaningless values and ensure internal consistency of the database. For this reason, it is advisable to plan the procedure of data entry and subsequent checks in advance.

Data entry procedures have changed over the recent years. Operators entering data into a computer manually are being replaced by computer technologies, while the universal distinction between three phases - (1) data collection, (2) data entry and (3) editing and checks - is becoming obsolete. These changes have largely been ushered by the emergence of computer-assisted data collection (CAPI, CATI, internet surveys etc.). Here, data entry occurs simultaneously with data collection and the software used includes a number of functions for checking data integrity. Thus, different techniques of checks are used that offer different correction options than in a classic survey design. Even where computers are not applied directly in the field, data can still be digitised with scanners. In spite of its limits (e.g. in recording textual answers), greater automation of processes generally prevents some types of errors, produces other types of errors, and changes the options for checking data.

The integrity of a data file is based on its structure and on links between data and elements of documentation such as variable labels and value labels. Below is a set of recommendations on minimising errors based on the data management guides of the UKDA [2010] and the ICPSR [2009: 13–14] and a book edited by Robert Groves [Groves et al. 2004: 319–321]:

- Manual data entry requires routine and concentration. Operators should not be burdened by multiple tasks. Tasks such as coding and data entry should be implemented separately.
- Final entry should be done through a smaller rather than a larger number of steps. This reduces the likelihood of errors.
- A great advantage lies in the use of specialised software with which it is possible to set the range of valid values for each category and to apply filters to manage the entry process (or the entire data collection process in the case of computer-assisted interviewing).[25] These automatic checks prevent meaningless values from being entered and often help to discover inconsistencies that arise when some values are skipped or omitted, and they make the interviewer's or operator's work substantially clearer and easier, thus generally reducing the number of errors they make.
- Data entry errors can usually be prevented if data entry is conducted twice and the results are compared. For example, double data entry is a standard for scanning.
- Check the completeness of records.

[25] If we do not have specialised software then we can use applications like MS Excel. Even they will allow us to reduce errors if we abide by a set of rules. For example, the University of Tennessee published instructions on data entry in MS Excel [SSC 2007].

- There are multiple methods for logical and consistency checks, including the following:
    - check the value range (e.g. a respondent over the age of 100 is unlikely),
    - check the lowest and highest values and extremes,
    - check the relations between associated variables (e.g. educational attainment should correspond with a minimum age, the total number of hours spent doing various activities should not exceed 100% of available time),
    - compare with historical data (e.g. check the number of household members with the previous wave of a panel survey).
- Many checks can be conducted automatically by computer. Even logical checks can be programmed directly into specialised CAPI, CATI or data entry software. Here, software can distinguish between permanent rules that cannot be bent and warnings that only notify the operator of entering an unlikely value.
- A certain percentage, e.g. 5–10% of all records, should be subject to a more detailed, in-depth check.
- Changes should be documented and original data should be restorable.

We can either delete or try to correct error values. Simple data entry errors can be easily corrected based on comparison with respondents' original answers. However, we should bear in mind that inconsistencies can also be generated by the respondents themselves, and a correction should make a minimum or no changes or reductions to their original answers. Any replacement of values must be planned and done in conformity with the concepts of measurement.

3.6. ANONYMISATION

The protection of respondents' personal data is one of the most important ethical and legal requirements of data management and personal data can be processed only in line with the respondents'informed consent (see above). The goal of social research is not to identify information about individuals, but to obtain generalised information. Consequently, it is often possible to avoid working with personal data in a research study. If the database is not anonymous and either the respondents'informed consent does not allow the personal data processing or the personal data are not neccessary for our research intentions, then the database has to be anonymised as soon as possible.

However, many databases that appear anonymous at first sight may actually harbour a significant risk of revealing respondents' identity. In quantitative research, this is particularly the case of surveys among smaller groups, surveys identifying some information in great detail, and ones that deal with certain specific cases in a file. Therefore, every data file should be assessed and analysed for the risk of disclosing respondents' identities before it can be considered anonymous. In some cases, methods of data anonymisation can ensure anonymity without critically damaging data quality.

A database is not anonymous if natural persons to whom the data in the database relate can be identified based on direct or indirect identifiers (see also point 2.3 above). Direct identifiers are, for example, names, national identification numbers, addresses, telephone numbers, respondents' photographs etc. Indirect identifiers make a person's identification possible when connected with other known data, for example, about a person's job, place of residence, workplace etc., or based on the extreme values of some variables. Indirect identification may also occur when several variables in a file are combined.

The following are basic anonymisation methods:

- Removing direct identifiers: Direct identifiers can often be replaced by anonymous codes while their basic functions are maintained. For example, the national identification number is removed and a unique questionnaire ID is kept which does not point to a specific person but helps distinguish between cases in the file.
- Removing or replacing interconnections with other available non-anonymous databases or sources of information.
- Aggregating information or reducing the variable's level of detail: Some information can be aggregated into categories referring to broader groups of subjects without losing informational value. For example, year of birth is recorded instead of the complete date of birth, or region of residence is recorded instead of the exact address. Special attention should be paid to geographic identifiers because persons can often be identified when the names of smaller municipalities are combined with other variables.

- Treating extreme values of variables: The risk of identifying persons based on atypical, extreme values can often be eliminated by introducing minimum and maximum limits of the range of valid values.

For example, in a panel survey we need to keep non-anonymous identifiers in order to build interconnections with data from other waves of the survey, but we do not need them for the analysis. Thus, we can create two files that can be interconnected by means of a unique ID. The first file contains all the survey information we need for the analysis, it is anonymous and can only be accessed for the purposes of investigation. The other file contains personal data, it is secured in terms of personal data protection, and is used exclusively for the purposes of building interconnections with data from the other waves.

3.7. WEIGHTING

Weights of sample survey data are constructed in order to take into account the characteristics of sampling design and correct identifiable deviations from population characteristics. Each individual case in the file is assigned a certain coefficient – individual weight – which is used to multiply the case in order to attain the desired characteristics of the sample. If the weight of a case equals 1 then the values measured are not adjusted.

Using the weights may be dysfunctional if some methods of analysis are employed. There are also general theoretical and methodological issues which discourage some researchers from using weights. Either way, the type and purpose of weighting cannot be omitted from the final decision. For this reason I here mention the fact that there are different types of weights for different purposes, though I do not devote any attention to specific procedures of weight construction:

- **Design weights** are constructed in order to mutually adjust individual units' probabilities of being sampled, which are normally not equal when complex sampling procedures combining multiple methods (stratification, group sampling) in several stages are implemented. For example, we want to adjust the probabilities of being sampled for all respondents in households. While individuals are the sampling units, households are sampled in the first stage. Therefore, respondents' probabilities of being selected depend on the number of household members.
- **Non-response weighting:** During the implementation of a survey, we are normally not able to get a response from some units sampled due to their refusal, our failure to contact them, or other administrative reasons. Response rates differ between various population groups and those inequalities can be compensated by weighting.
- **Post-stratification weighting:** This is done in order to achieve a distribution equal with that of some known characteristics of the population (e.g. sex, educational attainment).
- **Population size weighting:** Different groups may be represented in the database in different proportions than they are in reality. Such discrepancies are normally compensated through weighting. For example, international data files combine data from various countries. However, similarly large surveys are usually implemented in each of these countries, although their total populations are radically different in size. If we want to analyse data about large populations, such as in Europe, then we have to adjust the proportions in the representation of individual European countries.
- **Combined weighting:** Several different types of weights for different purposes may be constructed in the file. Subsequently, they are combined into a final, combined weight.

Table 6 provides an illustration of weighting in an example of an analysis of data from the European Social Survey (ESS). Two weights are constructed in the data files originating from the main questionnaires of the survey's individual waves. Design weighting takes into consideration different probabilities of being sampled given the sampling methods implemented in individual countries; population size weighting corrects the fact that the individual countries' sample sizes are very similar and the fact that there is a large variation in the size of their total populations.

TABLE 6: EXAMPLE OF WEIGHTING FOR DIFFERENT ANALYSES OF THE EUROPEAN SOCIAL SURVEY DATA FILE

	Example – voter turnout (% of respondents voting in the last election)	Weights to be used	
		Design weight	Population weight
To examine data from a single country – whether a single variable or a cross-tabulation	Voter turnout in Germany X	X	
	Voter turnout in Germany by age and gender	X	
To compare results for two or more countries separately – without using totals or averages	Compare voter turnout in France, Germany and the UK	X	
To combine countries – whether on a single variable or via a crosstabulation	Voter turnout in Scandinavia	X	X
	Voter turnout in the EU	X	X
	Voter turnout across all countries participating in the ESS	X	X
	Compare voter turnout between EU member states and accession countries	X	X
	Voter turnout by age group across all ESS participating countries	X	X

Source: European Social Survey [ESS 2008b: 2].

3.8. DATA FILE DOCUMENTATION

While documentation generally facilitates the utilisation of data, it represents a vital condition for meaningful long-term preservation of the data file and for secondary analysis. For this purpose, it is important to equip the data file with information on how the data came to exist, what it means, what content and structure it has, and what adjustments have been made to it. Moreover, the development of the documentation constitutes also the first phase of data interpretation and may significantly influence the formation of research results. For these reasons, the quality of documentation is also one of the components of data quality and the documentation requirement along with the sets of minimum information about the survey that have to accompany survey data are prescribed by general research standards (e.g. ISO 20252:2006, ESOMAR [2008], AAPOR [1997]).

The field of archiving and data management uses the term metadata to refer to documentation. Metadata, i.e. data on data, is an integral part of the data file. Documentation includes a set of items describing the nature of the project. The contents, structure, way of acquisition of information and format of documentation should be planned at the very start of the research process because a lot of relevant information has to be recorded in different stages of the survey and would be costly or impossible to collect in retrospect. For example, in order to document the different types of non-response (see Table 7), we need to make sure such information is gathered and recorded for both successful and unsuccessful interviews during data collection; also, consistency checks and edits in the data file are more easily documented during their implementation.

The format of documentation should correspond with the intended uses of the data. We have to take into consideration our own intentions, the requirements of the recipients of our research results, or the requirements of archives that will be making our data available. Selected information, especially documentation of variables, is usually integrated in the data file. Other metadata should be included in a special structured document. The wording of questions might also be included in it, although this is often also provided in a separately attached research instrument. Even if variable documentation forms part of the data file, some variables should be accompanied by additional specific information, for example weighting algorithms, a detailed description of performed transformations, the syntax used in converting classifications etc.

In the past, data was usually accompanied by so-called codebooks or other reference guides printed on paper. These contained information about the research study and its methods, lists of codes, or even lists of frequencies and selected contingency tables. At present, such documentation tends to be provided electronically, which makes it easier to search through and connect with data and other materials. Like data, ‘metadata’ also have to be in line with the objectives of long-term preservation, i.e. their formats should be stable vis-à-vis software development. The DDI internati-

onal standard covers the format and contents of documentation in a comprehensive way. It presents a universal structure of documentation for the widest range of types of social science data and purposes, including long-term preservation of metadata and recording of data file history.

TABLE 7: EXAMPLE OF INDICATORS REQUIRED FOR THE DOCUMENTATION AND ANALYSIS OF NONRESPONSE IN THE INTERNATIONAL SOCIAL SURVEY PROGRAMME (ISSP) AND SELECTED RESPONSE INDICATORS. THE DATA COME FROM THE CZECH SURVEY OF ISSP 2009 ON SOCIAL INEQUALITY IV

A.	Total number of starting or issued names/addresses (gross sample size)	2000
B.	Addresses which could not be traced at all/selected respondents who could not be traced	0
C.	Addresses established as empty, demolished or containing no private dwellings	14
D.	Selected respondent too sick/incapacitated to participate	43
E.	Selected respondent away during survey period	139
F.	Selected respondent had inadequate understanding of language of survey	3
G.	No contact at selected address	61
H.	No contact with selected person	14
I.	Personal refusal at selected address	212
J.	Proxy refusal (on behalf of selected respondent)	277
K.	Other refusal at selected address	32
L.	Other type of unproductive reaction(please describe in full details)	0
M.	Full productive interview (net sample size)	1205
N.	Partial productive interview	0
Response Rate = M / (A – (B + C))		60,7 %
Co-operation Rate = M / (M + I + J + K)		69,8 %
Contact Rate=M / (M + G + H)		94,1 %

Source: ISSP on Social Inequality IV, Czech Republic (Institute of Sociology ASCR, Factum).

Here I will focus on the basic extent of documentation recommended for sample survey data assuming their usage for secondary analysis.

When documenting data files from sample surveys, we should distinguish two general levels – information about the survey and information about the data. In particular, it is important not to omit the following items:

(a) Information about the survey

- Data file origin: the title of the survey (including acronyms and their explanation, alternative titles in foreign languages etc.), institutional information (authors, implementing, funding and commissioning institutions, grant numbers etc.), project abstract, objectives, concepts, hypotheses, and references to follow-up projects.
- Description and methods of data collection: description of all the sources the data originate from (e.g. for derived data, data added from other sources), the time period of data collection, temporal and geographical coverage, target population, units of observation, description of sampling design including frame, methods of data collection, the wording of the questions in the questionnaire, the original research instrument and other materials used in data collection (letters of invitation, instructions for interviewers etc.), classification schemes and concepts applied, response rate and other assessments (e.g. known deviations from the research population), identification of methodological changes for time series and longitudinal studies.
- Description of data files: specification of the version and the edition of a collection, the structure of the data files, a specification of associations and interconnections (including technical information for forming links between files etc.), size information (the number of units and variables), information about formats and compatibility.
- Data edits and modifications: methods and results of integrity checks, validation, data cleaning, or any other applicable procedures for increasing data quality (calibration, the imputation of missing values, checks and corrections of transcripts etc.), anonymisation, transformation and construction of derived variables, weighting (the identification of variables for weighting and a description of weighting methods and design).
- Access to data: a definition of authorised persons, a specification of terms of use, information about personal data protection.
- Cataloguing and citation information: bibliographic information, suggested citation, key words, cataloguing information.
- References to related materials and sources if applicable.

(b) Information about the data

- Information about the variables in the file: the names, labels and descriptions of variables, their values, a description of derived variables or, if applicable, frequencies, basic contingencies etc. The exact original wording of the question should also be available.
- Information about the cases in the file: a specification of cases if applicable.

3.9. VERSIONS AND EDITIONS, ENSURING AUTHENTICITY

The above-described data management procedures typically result in several versions of the data file. New versions are created when errors that occur after data cleanup during data analysis are subsequently corrected; when the data are processed for the purpose of analysis; or when new data or data from other sources are added. The treatment of errors and the inclusion of new data may result in the publication of different editions of the same dataset which may even differ substantially in their contents (e.g. when data from additional countries are included in an international database). However, different working versions of the file are even created in research studies with simple databases or for the researchers' own needs. Thus, it is advisable to keep track of the contents of each version and avoid overwriting the authentic original file.

A good strategy for managing data file versions and editions is necessary to ensure that the data are safe, the data file contents are comprehensible, and mismatches and mistakes are avoided. The objective is (1) to clearly distinguish between individual versions and editions and keep track of their differences, (2) to ensure data authenticity, i.e. prevent unauthorised modification of files and loss of information. The following basic rules apply [see also UKDA 2010: Version Control & Authenticity]:

- Establish the terms and conditions of data use and make then known to team members and other users.
- Distinguish between versions shared by multiple researchers and individuals' working versions.
- Introduce clear and systematic naming of data file versions and editions.
- Maintain records about the creation of versions and editions, their specific contents and mutual relations.
- Document any changes made.
- Keep original versions of data files, or keep documentation that allows the reconstruction of original files.
- Create a 'master file' and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes.
- If there are several copies of the same version, check that they are identical.

3.10. DATA PRESERVATION: BACKUPS, FORMATS, MEDIA

Digital media are unreliable in principle. Software development results in frequent format changes, which affects compatibility. Institutions preserving data go through organisational changes, too. Additional risks are related to software failure and viruses, inadequate human intervention, and natural disasters. As a result, data preservation represents a process, rather than a state, and requires a well-considered approach.

Backup is the fundamental element of security. The extent and methods of backup implementation should be in line with its objective, namely a complete restoration of lost files. Backup should take place in a systematic and regular fashion. For that purpose, many institutions have formulated uniform backup policies. Such policies should take into account the needs related to research data files. Moreover, they should cover, if applicable, any substantial modifications to data as soon as possible after they are made. Backup processes can be automated by means of specialised software applications and backup devices. A backup copy has to be labelled and documented. Data should be deposited in such formats and on such media that are favourable for long-term preservation (see below). Its completeness and integrity should be verified and checked. Given the risks backup is expected to cover (e.g. floods, fire), backup copies should be located elsewhere than the original data.

In order to secure our data and ensure its long-term preservation, we should choose adequate formats and a favourable location.

(a) Data formats and documentation

Software and software formats are developing rapidly. For the reasons of short-term operability, it is advisable to choose a format associated with specific software applications. We also have to take into consideration how widespread this format is and to what extent the computer environment is friendly to it in terms of the compatibility and availability of conversion tools. We should keep in mind that veryspecific formats (e.g. SPSS files with the *.sav extension, STATA files with the *.dta extension, MS Access files with the *.mdb or *.accdb extensions, Dbase files with the *.dbb extension) undergo frequent changes and may not be fully transferable, even between different versions of the same software application. However, some software utilises the so-called portable versions of formats that are associated with a specific application or group of applications and allow easy transfer of data between different versions and hardware platforms (e.g. ‘portable’ SPSS files with the *.por extension or ‘transport’ SAS files). Open, widespread formats are more advisable for long-term storage as they typically undergo fewer changes.

The use of national specific characters in the database, e.g. Czech diacritics, creates another issue. In this case we need to pay great attention to character coding. Some character encoding systems

(e.g. Microsoft Windows - ANSI character encodings) do not cover all characters in one system. As a result, the appropritae language environment (Central European languages) has to be set to ensure correct display, and this cannot always be done. Other coding systems (e.g. UTF 8 and higher) allow several character sets to be correctly displayed simultaneously.

Long-term preservation of quantitative data (depending on the arrangement and type of data) is normally best done using simple text (ASCII) formats and a structured documentation file with information about the variables in it, their position in the file, formats, variable labels, value labels etc. Records for each unit in such a file are normally located on separate rows. If any of the records extend over multiple rows then we should note this in the documentation. In terms of the location of variables in the file, a distinction is made between fixed and free formats. In a fixed format, variables are arranged in columns and their exact positions, i.e. the start and end of each variable, are known.

The position of a variable is irrelevant in a free format – the data for each variable is separated by blanks or specific characters, such as a tab space or a dash. For instance, in the Czech language environment it is necessary to remember that the comma, a frequent data separator in English versions of databases, is used as a decimal separator instead of a dot. It is essential tomake sure that there is not more than one possible meaning to a symbol in a file.

Digital versions of paper documentation are usually kept in PDF/A format. This is the official version of the PDF format for archiving conforming to the ISO 19005-1:2005 standard. It guarantees independence from the platform, includes all display information (including fonts, colours etc.) and metadata in the XML format, and disallows encryption or password protection. Structured textual documentation should be saved in a simple text format, with tags and in line with a standard structure (e.g. DDI).

(b) Digital media

The choice of media for the long-term preservation of data depends not only on the type of media but also on the quality of different media of the same type. It also importantly depends on the methods of storage and, last but not least, on technological changes – after several years, compatible readers are difficult to find for some media. For example, a high-quality 5.25 inch floppy disc may preserve data for up to twenty years, but it is easily damaged if handled without care. At the same time, very few institutions these days have the equipment for reading this once widespread medium.

Generally speaking, all digital media are subject to high risk of damage, loss of information, and outdating due to technological development. Therefore, it is of primary importance to formulate a strategy for data backup and preservation in consideration of existing risks and expected developments. The

UK Data Archive [UKDA 2010: Storing your data] recommends regular copying of data to new media. Data preserved on optical media (which are highly susceptible to physical damage and sensitive to changes in temperature, air humidity and light conditions) should be copied every two to five years. The situation is similar for magnetic media, which are susceptible to physical damage. Therefore, a reasonable strategy of data preservation envisions at least two different methods of data storage. The selected methods should take into account the risks of damage to digital media.

3.11. ARCHIVING

The objectives of long-term preservation and continuous distribution of data for the purposes of secondary analysis can be effectively achieved by a publicly accessible social science data archive (see Box 8 for selected social science data archives and their Web addresses). Such an archive, however, may not be able or willing to accept every kind of data. The general goal is to make as much data as possible available in a user-friendly and useful format. Therefore, the main conditions for archiving are as follows:

- relevance for the archive's orientation,
- data readability and documentation (formats and media),
- authorisation of access in line with copyright law,
- treatment of data in terms of personal data protection,
- completeness of data and documentation so that the data can be understood by users outside the research team.

Archives tend to search for and acquire data actively, and make substantial efforts to process data for archiving. Given the requirements some funding agencies stipulate for supporting project proposals,[26] many archives may place high demands on data depositing and may transfer a substantial proportion of the costs of preparing the data for archiving to researchers. Consequently, if you are planning to deposit your data in an archive, you should acquaint yourself with its terms and conditions in advance.

[26] Many important funding agencies (e.g. the US NSF, the UK Research Councils and several others) require the data to be archived and made available as a condition for a research project to be regarded as completed.

BOX 8: SELECTED SOCIAL SCIENCE DATA ARCHIVES AND ORGANISATIONS

Country	Organisation	URL
International	CESSDA – Consortium of European Social Science Data Archives	http://www.cessda.org
United States/ International	ICPSR – Inter-university Consortium for Political and Social Research	http://www.icpsr.umich.edu
International	IFDO – International Federation of Data Organisations for Social Science	http://www.ifdo.org/
Poland	ADP – Polish Social Data Archive	http://www.ads.org.pl/
Russia	DBSR – Data Bank of Social Research at the Institute of Sociology Russian Academy of Sciences	http://www.isras.ru/ Databank.html
Lithuania	LiDA – Lithuanian Data Archive for Social Sciences and Humanities	http://www.lidata.eu/
Czech Republic	CSDA – Czech Social Science Data Archive	http://archiv.soc.cas.cz/
Austria	WISDOM – Austrian Data ArchivE	http://www.wisdom.at/
Denmark	DDA – Danish Data Archive	http://samfund.dda.dk
Estonia	ESSDA – Estonian Social Science Data Archive	http://psych.ut.ee/esta/
Finland	FSD – Finnish Social Science Data Archive	http://www.fsd.uta.fi
France	Réseau Quetelet	http://www.reseau-quetelet.cnrs.fr/spip/
Germany	GESIS – Leibniz Institute for the Social Sciences	http://www.gesis.org
Greece	Greek Social Data Bank (GSDB)	http://www.gsdb.gr/
Hungary	TÁRKI Data Archive	http://www.tarki.hu/en/services/da/
Ireland	ISSDA – Irish Social Science Data Archive	http://www.ucd.ie/issda/
Israel	ISDC – Israel Social Science Data Center	http://isdc.huji.ac.il/
Italy	ADPSS – Data Archive for Social Sciences	http://www.sociologiadip.unimib.it/sociodata/
Japan	SSJDA – Social Science Japan Data Archive	http://ssjda.iss.u-tokyo.ac.jp/en/
Luxembourg	CEPS/INSTEAD	http://www.ceps.lu/
Netherlands	DANS – Data Archiving and Networked Services	http://www.dans.knaw.nl/
Norway	NSD – Norwegian Social Science Data Services	http://www.nsd.uib.no
Portugal	APIS – Portuguese Social Information Archive	http://www.apis.ics.ul.pt
Romania	RODA – Romanian Social Data Archive	http://www.roda.ro/
Slovakia	SASD – Slovak Archive of Social Data	http://sasd.sav.sk/sk/
Slovenia	DP – Social Sciences Data Archive	http://www.adp.fdv.uni-lj.si/
South Africa	SADA – South African Data Archive	http://sada.nrf.ac.za/
Spain	CIS Databank	http://datosbd.cis.es
Sweden	SNDS – Swedish National Data Service	http://snd.gu.se
Switzerland	FORS DARIS – FORS Data and Research Information Services	http://www2.unil.ch/daris
United Kingdom	UK Data Archive	http://www.data-archive.ac.uk/

Source: European Social Survey [ESS 2008b: 2].

CONCLUDING REMARKS

Good data management is important in order for research work to run smoothly and to produce high--quality scientific results. At the same time, it represents a fundamental condition of data sharing. Data sharing is one of the basic trends in the contemporary organisation of collaborative social research and in the methods of scientific work which are based on using and combining empirical evidence from different sources. Data sharing is becoming a standard and often a duty, too. Ultimately, it is becoming a condition for attaining international excellence in research.

Data archives and various publicly accessible repositories specialising in data management represent the means of data sharing. However, primary responsibility for the quality of databases remains with the researchers and producers of data. The information provided in this paper should help them approach data management systematically, tackle its tasks effectively, and avoid some problems.

REFERENCES

AAPOR. 1997. *Best Practices for Survey Research*. Deerfield: American Association for Public Opinion Research (AAPOR). On-line: http://www.aapor.org/Best_Practices1.htm [visited 12 September 2013].

Czech Republic. 2011. *Act No. 101/2000 Coll., Consolidated version of the Personal Data Protection Act*. Published on the website of the Office for Personal Data Protection. On-line: [visited 12 September 2013].

Czech Republic. 2006. *Consolidated version of Act No. 121/2000 on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), as amended by Act No. 81/2005, Act No. 61/2006 and Act No. 216/2006*. English translation of the Act published on the Web of the Ministry of Culture of the Czech Republic. On-line: http://www.mkcr.cz/assets/autorske-pravo/12--AZ_2006_v_AJ.pdf

CESSDA. 2009. *Sharing Data*. Online presentation. Bergen: CESSDA. On-line: http://www.cessda.org/sharing/ [visited 12 September 2013].

ESOMAR. 2008. ICC/ESOMAR *International Code of Marketing and Social Research Practice*. ESOMAR World Research Codes and Guidelines. Amsterdam: ESOMAR. On-line: http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_ICC-ESOMAR_Code_English.pdf [visited 12 September 2013].

ESRC. 2010. ESRC *Research Data Policy*. Swindon: Economic and Social Research Council (ESRC). On-line: http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf [visited 12 September 2013].

ESS. 2008a. ESS 2008 *Data Protocol*. Edition 1.2. Bergen: Norwegian Social Science Data Services (NSD). On-line: http://www.europeansocialsurvey.org/docs/round4/survey/ESS4_data_protocol_e01_2.pdf [visited 12 September 2013].

ESS. 2008b. *Weighting European Social Survey Data*. Bergen: Norwegian Social Science Data Services (NSD). On-line: http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data.pdf [visited 12 September 2013].

European Commission. 2007. *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation*. Brussels: Commission of the European Communities. On-line: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0056:FIN:EN:PDF [visited 12 September 2013].

European Commission. 2012a. *Commission Recommendation of 17 July 2012 on access to and preservation of scientific information (2012/417/EU)*. Brussels: Commission of the European Communities. On-line: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:EN:-PDF [visited 12 September 2013].

European Commission. 2012b. *Communication from the Commission to the European Parliament, the Council, the European Economic And Social Committee and the Committee of the Regions. Towards better access to scientific information: Boosting the benefits of public investments in research*. On-line: http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf [visited 12 September 2013].

Council of the European Union. 2007. *Conclusions on Scientific Information in the Digital Age: Access, Dissemination and Preservation*. 2832nd Competitiveness (Internal market, Industry and Research) Council meeting, Brussels, 22 and 23 November 2007. Brussels: Council of the European Union. On-line: http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf [visited 12 September 2013].

Eynden, Veerle Van den, Libby Bishop, Laurence Horton, Louise Corti 2010. *Data Management Practices in the Social Sciences*. Colchester: UK Data Archive (UKDA). On-line: http://www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf [visited 12 September 2013].

Eynden, Veerle Van den, Louise Corti, Matthew Woolard, Libby Bishop, Laurence Horton 2011. *Managing and Sharing Data. Best Practice for Researchers*. Colchester: UK Data Archive (UKDA). On-line: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf [visited 12 September 2013].

Freed-Taylor, Marcia 1994. "Ethical considerations in European cross-national research." *International Social Science Journal* 142: 523–532.

Ganzeboom, Harry B. G. 2010. *Do's and Don'ts of Occupation Coding with an Extension to ISCO-08*. Working paper of the Department of Social Research Methodology. Amsterdam: Free University of Amsterdam. On-line: http://home.fsw.vu.nl/hbg.ganzeboom/isco08/..%5Cpdf%5C2010-do-and--donts-occupation-coding-%28paper-version4%29.pdf [visited 12 September 2013].

Green, Ann G., and Myron P. Gutmann 2007. "Building Partnerships among Social Science Researchers, Institution-based Repositories, and Domain Specific Data Archives." *OCLC Systems & Services* 23 (1): 35–53. On-line: http://dx.doi.org/10.1108/10650750710720757 (DOI – permanent URL).

Groves, Robert M., Floyd J. Fowler, Mick P. Couper and James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken: John Wiley & Sons.

Humprey, Charles 2006. *e-Science and the Life Cycle of Research*. Unpublished electronic document. On-line: http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc [visited 12 September 2013].

ICO. 2013. *Proposed new EU General Data Protection Regulation: Article-by-article analysis paper.* Wilmslow: Information Commission's Office (ICO). On-line: http://www.ico.org.uk/news/~/media/documents/library/Data_Protection/Research_and_reports/ico_proposed_dp_regulation_analysis_paper_20130212_pdf.ashx [visited 11 September 2013].

ICPSR. 2009. *Guide to Social Science Data Preparation and Archiving. Best Practice Throughout the Data Life Cycle*. 4th Edition. Ann Arbor: Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan.

ICPSR. 2011. *Guidelines for Effective Data Management Plans*. Ann Arbor: Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan. On-line: http://www.icpsr.umich.edu/files/ICPSR/dmp/DataManagementPlans-All.pdf [visited 12 September 2013].

ICPSR. 2012. *Guide to Social Science Data Preparation and Archiving. Best Practice Throughout the Data Life Cycle*. 5th Edition. Ann Arbor: Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan. On-line: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf [visited 12 September 2013].

Iverson, Jeremy 2009. „Metadata-Driven Survey Design". *IASSIST Quarterly 2009* (Spring-Summer): 7–9. On-line: http://www.iassistdata.org/downloads/iqvol3312iverson.pdf [visited 12 September 2013].

JISC, TLTP. 1998. *JISC/TLTP Copyright Guidelines*. Bristol: Joint Information Systems Committee (JISC), Teaching and Learning Technology Programme (TLTP).

Krejčí, Jindřich, Yana Leontiyeva (eds.). (2012) *Cesty k datům. Zdroje a management sociálněvědních dat v České republice* [Pathways to Data. Sources and Management of Social Science Data in the Czech Republic]. Prague: SLON.

NSF. 2006. *Grant General Conditions*. GC-1, March 15, 2006. Arlington: National Science Foundation (NSF).

OECD. 2004. *Declaration on Access to Research Data from Public Funding*. Annex 1 of the Final Communiqué from the Meeting the OECD Committee for Scientific and Technological Policy at Ministerial Level, January 29–30, 2004. Paris: Organisation for Economic Co-operation and Development (OECD). On-line: http://www.oecd.org/document/15/0,2340, en_21571361_21590465_25998799_1_1_1_1,00.html [visited 12 September 2013].

OECD. 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: Organisation for Economic Co-operation and Development (OECD). On-line: http://www.oecd.org/dataoecd/9/61/38500813.pdf [visited 12 September 2013].

Ruusalepp, Raivo 2008. A *Comparative Study of International Approaches to Enabling the Sharing of Research Data*. London: Joint Information Systems Committee (JISC)/Data Curation Centre (DCC), 2008. On-line: http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-Sharing-Report.pdf [visited 12 September 2013].

SSC. 2007. *How to Use Excel for Data Entry*. Statistical Consulting Centre (SCC) University of Tennessee. Knoxville: University of Tennessee. On-line: http://oit.utk.edu/scc/HowToUseExcelForDataEntry.pdf (visited 27 June 2011)

UKDA. 2010. Create & Manage Data. Online presentation. Colchester: UK Data Archive (UKDA). On-line: http://www.data-archive.ac.uk/create-manage [visited 12 September 2013].

UKDA. 2011. *UK Data Service – Data management costing tool and checklist*. Colchester: UK Data Archive (UKDA). On-line: http://data-archive.ac.uk/media/247429/costingtool.pdf [visited 12 September 2013].

WAPOR undated. *WAPOR Code of Ethics*. Lincoln: World Association for Public Opinion Research (WAPOR). On-line: http://wapor.unl.edu/wapor-code-of-ethics/ [visited 12 September 2013].

SUMMARY

The quality of data management affects the quality of data and importantly determines the adequacy of research findings. Systematic data management is, above all, a necessary condition for preventing errors and false findings, but it can also save a lot of time and make research work both clearer and easier. Higher demands are placed on data management especially when we plan the long-term preservation and sharing of data between different research teams. In recent years, the requirements of archiving and providing open access to social science data for the purpose of secondary data analysis have become highly important parts of scientific work. Consequently, demands on professional data management have also risen.

The paper opens with a discussion of the conceptual background to data management in contemporary social research, which is shaped by efforts to implement open access policies and a cyclical view of the life of data. It then briefly reviews the principles of data management in the preparation stage of the research process. Finally, it guides the reader through the different areas of data management during the research process.

The exchange of scientific information and results is crucial for the development of research today and this is equally true for the sharing of social science data. Data are often produced at considerable expense for the public resources. Therefore, it is logical to ask researchers who are recipients of public funding to make further use of their databases possible. The OECD codified a set of basic principles and guidelines for access to research data resulting from public funding (2007). Current practices of organisation of social science research based on data sharing are gradually institutionalised on the level of international and national science policies.

The requirements of data sharing have changed the functions of data management. When a dataset is being produced, one must count on archiving and publicising it without knowing by whom and for what purposes the data are going to be used. Like in other disciplines, research in the social sciences takes the form of a cycle where the results of one research study feed back into the research process as background for other research studies. In an environment characterised by open access to data, secondary analysis of research data plays an important role in this cycle. This gives data a new function in the dissemination and reproduction of knowledge and this function must be reflected in the ways data are managed.

Every proposal for empirical research should start not only by reviewing the relevant literature but also by carefully reviewing available data sources. This is equally important for studies that are based on their own data collection, because besides data one can also utilise the accompanying information about methodologies, procedures and research instruments from prior research.

The main ethical requirements of data management consider (1) protection of respondents from the potential harmful effects of research whenever the data are worked with, archived, made available, or made subject to secondary analysis; (2) the respects to informed consent and decisions of survey respondents; and (3) requirement of adequate utilisation of the data in respect to the efforts respondents made to participate in the research study, public investments and nature of the data. The issues of personal data protection should be given adequate attention as early as the stage in which a research proposal is drafted. To underestimate them would not only constitute a violation of research ethics but might also restrict or completely prevent the researchers' intentions from being fulfilled and in particular the data from being made available for secondary research.

Social scientists often pay too little attention to and have ample misconceptions about copyright issues when it comes to databases and data sharing. Copyright and intellectual property protection is complicated and a thorough treatment of these issues requires professional legal advice. In each research institution such legal advice should result in the creation of ground rules and standard practices for its employees to follow. Nevertheless, each researcher should be aware of at least the basic rules, thus this paper provides an overview of such basic rules.

Data files may have different internal structures and a research study may encompass several data files in different relations to one another. The following basic types of data files are distinguished in terms of structure: flat file, hierarchical file and relational database. The file's structure is further shaped by the arrangement and labelling of variables. There are relationships between variables, e.g., an original and a derived variable or sets of variables linked to the measurement of the same phenomenon. At the same time, there are links to other elements of the research study such as the questionnaire, data source, or another dataset.

There are different types of variable values. The fact of numeric coding is shared by all statistical software applications and, among other things, this facilitates data conversion and measurement comparisons. Our decisions about the specific structure of coded categories also affect the prevention of errors in data entry and data processing and define the basis of analysis. Generally speaking, coded categories should refer to the contents of the hypotheses tested. At the same time, one must take into consideration any possible uses of the data in future and maximise the data's informational value. Not all the questions in a questionnaire are answered by all respondents, which results in missing values for the corresponding variables. It is advisable to distinguish between various missing data situations. The integrity of a data file is based on its structure and on links between data and elements of documentation such as variable labels and value labels.

If the database is not anonymous and either the respondents'informed consent does not allow the personal data processing or the personal data are not neccessary for our research intentions, then the database has to be anonymised. The basic anonymisation methods are following: removing direct identifiers, removing or replacing interconnections with other available non-anonymous databases or sources of information, aggregating information or reducing the variable's level of detail and treating extreme values.

Weights of sample survey data are constructed in order to take into account the characteristics of sampling design and correct identifiable deviations from population characteristics.

While documentation generally facilitates the utilisation of data, it represents a vital condition for meaningful long-term preservation of the data file and for secondary analysis. Moreover, the development of the documentation constitutes also the first phase of data interpretation and may significantly influence the formation of research results. When documenting data files from sample surveys, we should distinguish two general levels – information about the survey and information about the data.

The data management procedures typically result in several versions of the data file. A good strategy for managing data file versions and editions is necessary to ensure that the data are safe, the data file contents are comprehensible, and mismatches and mistakes are avoided. The objective is (1) to clearly distinguish between individual versions and editions and keep track of their differences, (2) to ensure data authenticity, i.e. prevent unauthorised modification of files and loss of information.

Digital media are unreliable in principle. Software development results in frequent format changes, which affects compatibility. Institutions preserving data go through organisational changes, too. Additional risks are related to software failure and viruses, inadequate human intervention, and natural disasters. As a result, data preservation represents a process, rather than a state, and requires a well-considered approach. Backup is the fundamental element of security.

Introduction to the Management of Social Survey Data
Jindřich Krejčí

Published by the Institute of Sociology of the Czech Academy of Sciences, Prague 2014.
First edition.
Print run: 250 copies.
ISBN 978-80-7330-252-8

Edited by Marta Svobodová.
Translated by Jan Morávek.
Language editor: Robin Cassling

Cover picture by Jaroslav Kašpar.
Design, typeset and layout by Jaroslav Kašpar.
Printed by ERMAT Praha, s.r.o., Antala Staška 1021/55, Praha 4.

Address of publisher:
Institute of Sociology, CAS, Jilská 1, 110 00 Praha 1

Sales:
Press and Publications Department
Institute of Sociology, CAS, Jilská 1, 110 00 Praha 1
Tel: +420 210 310 217
prodej@soc.cas.cz

www.soc.cas.cz

The quality of data management affects the quality of data and importantly determines the adequacy of research findings. Systematic data management is, above all, a necessary condition for preventing errors and false findings, but it can also save a lot of time and make research work both clearer and easier. Higher demands are placed on data management especially when we plan the long-term preservation and sharing of data between different research teams. In recent years, the requirements of archiving and providing open access to social science data for the purpose of secondary data analysis have become highly important parts of scientific work. Consequently, demands on professional data management have also risen.

The paper opens with a discussion of the conceptual background to data management in contemporary social research, which is shaped by efforts to implement open access policies and a cyclical view of the life of data. It then briefly reviews the principles of data management in planning and designing the research project. Finally, it guides the reader through the different areas of data management during the research process.

ISBN 978-80-7330-232-0

SOÚ
Institute of Sociology CAS