



cessda **DMEG**

Data Management
Expert Guide

cessda.eu/DMEG

Offline version - January 2020



Contents

Introduction.....	5
Seven chapters.....	6
Chapter 1	7
Main take-aways	8
1.1 Benefits of Data management.....	9
1.2 Research data	13
1.3 Data in the social sciences.....	15
1.4 FAIR data	18
1.5 European diversity.....	20
1.6 Adapt your DMP: Part 1	28
Sources and further reading	31
Chapter 2	32
Main take-aways	33
2.1 Designing a data file structure.....	34
2.1.1 Organisation of variables	39
2.2 File naming and folder structure	43
2.3 Documentation and metadata	47
2.4 Adapt your DMP: part 2	57
Sources and further reading	58
Chapter 3	59
Main take-aways	60
3.1 Data entry and integrity.....	61
3.2 Quantitative coding.....	66
3.3 Qualitative coding.....	70
3.4 Weights of survey data	72
3.5 File formats and data conversion.....	76
3.6 Data authenticity	79
3.7 Wrap up: Data quality	82
3.8 Adapt your DMP: part 3	84
Sources and further reading	85

Chapter 4	86
Main take-aways	87
4.1 Storage.....	88
4.2 Backup	93
4.3 Security	97
4.4 Adapt your DMP: part 4	100
Sources and further reading	102
Chapter 5	103
Main take-aways	104
5.1 Ethics and data protection	105
5.2 Ethical review process.....	107
5.3 Processing personal data	113
5.3.1 Diversity in data protection.....	120
5.4 Informed consent.....	125
5.5 Anonymisation	133
5.6 Copyright	140
5.6.1 Diversity in copyright	143
5.7 Adapt your DMP: part 5	153
Sources and further reading	154
Chapter 6	155
Main take-aways	156
6.1 Towards archiving & publication	157
6.2 Selecting data for publication	159
6.3 Data publishing routes	160
6.4 Publishing with CESSDA archives	164
6.4.1 Citing your data	165
6.4.2 Licensing your data	167
6.4.3 Access categories	169
6.5 Promoting your data	172
6.6 Adapt your DMP: part 6	174
Sources and further reading	175

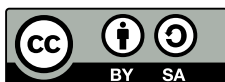
Chapter 7	176
Main take-aways	177
7.1 The process of data discovery	178
7.2 Data repositories as data resources	193
7.3 Resources for social media data.....	198
7.4 Access, use and cite data	200
7.5 Adapt your DMP: part 7	206
Sources and further reading	207
Contributors	208

CITATION

CESSDA Training Team (2017 - 2019). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. DOI: 10.5281/zenodo.3820473

Retrieved from <https://www.cessda.eu/DMGuide>

LICENCE



The Data Management Expert Guide by CESSDA ERIC is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. All material under this licence can be freely used, as long as CESSDA ERIC is credited as the author.

Introduction

This guide is designed by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (FAIR).

You will be guided by different European experts who are - on a daily basis - busy ensuring long-term access to valuable social science datasets, available for discovery and reuse at one of the CESSDA social science data archives.

Target audience and mission

This guide is written for social science researchers who are in an early stage of practising research data management. With this guide, CESSDA wants to contribute to professionalism in data management and increase the value of research data.

Overview

If you follow the guide, you will travel through the research data lifecycle from planning, organising, documenting, processing, storing and protecting your data to sharing and publishing them. Taking the whole roundtrip will take you approximately 15 hours, however you can also hop on and off at any time.

During your travels, you will come across the following recurring topics:

Adapt your DMP

As the data management plan (DMP) is an important tool to structure the research data management of your project, it plays a central role in this guide. Each chapter ends with a section with questions that are generally to be answered in a DMP. In the chapter's paragraphs you will be presented with the information you need to answer the proposed questions.

We have designed a list of DMP-questions especially for this Data Management Expert Guide. You can view and download the checklist as pdf (CESSDA, 2018a) or editable form (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

European diversity

Here you will learn about European diversity in data management. E.g., think of diversity in:

- » Funder requirements in making your Data Management Plan
- » Data protection law
- » Informed consent
- » Ethical review
- » The requirements of local data repositories

Expert tips



In each chapter, you will find advice from our experts. These expert tips aim to provide you with food for thought and practical how-to information.

Examples for quantitative and qualitative data

To give you in-depth tips on how to deal with your own specific type of data, chapters may contain examples of how to handle quantitative or qualitative data.

Seven chapters

The following seven chapters are currently available.

[Plan](#)

In this introductory tour, you will become aware of what data management and a data management plan (DMP) are and why they are important. General concepts such as social science data and FAIR data will be explained. Based on our recommendations and good practice examples, you will be able to start writing your DMP.

[Organise & Document](#)

If you are looking for good practices in designing an appropriate data file structure, naming, documenting and organising your data files within suitable folder structures, this chapter is for you.

[Process](#)

You will get acquainted with the topics of data entry and coding as the first steps of proper data management. To maintain the integrity of your data we will guide you in choosing the appropriate file format. You will also find out about version and edition management.

[Store](#)

To be able to plan a storage and backup strategy, you will learn about different storage and backup solutions and their advantages and disadvantages. Also, measures to protect your data from unauthorised access with strong passwords and encryption will be explained.

[Protect](#)

This chapter highlights your legal and ethical obligations and shows how a combination of gaining consent, anonymising data, gaining clarity over who owns the copyright to your data and controlling access can enable the ethical and legal sharing of data.

[Archive & Publish](#)

When you arrive at this chapter you will have learnt to differentiate between currently available data publication services. You will also find a number of stepping stones on how to promote your data.

[Discover](#)

How can you discover and reuse existing or previously collected datasets?

Chapter 1

Plan

Contents

Main take-aways	8
1.1 Benefits of Data management.....	9
1.2 Research data	13
1.3 Data in the social sciences.....	15
1.4 FAIR data	18
1.5 European diversity.....	20
1.6 Adapt your DMP: Part 1	28
Sources and further reading	31

[View the online version of this chapter](#)

Main authors of this chapter

Ulf Jakobsson, Swedish National Data Service (SND)

Ricarda Braukmann, Data Archiving and Networked Services (DANS)

Malin Lundgren, Swedish National Data Service (SND)

Introduction



This introductory chapter features a brief introduction to research data management and data management planning.

Before we get you started on making your own Data Management Plan (DMP), we will guide you through the basic concepts that you will need to understand beforehand. Research data, social science data, and FAIR data are some of the concepts you will discover.

Main take-aways

After completing this chapter you should be:

- » Familiar with concepts such as (sensitive) personal data and FAIR principles;
- » Aware of what data management and a data management plan (DMP) are and why they are important;
- » Familiar with the content elements that make up a DMP;
- » Able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt them to your own DMP.

1.1 Benefits of Data management

Research data management refers to how you handle, organise, and structure your research data throughout the research process. Data management:

- » Begins with your initial considerations regarding what will be necessary for using or collecting your particular type of data;
- » Includes measures for maintaining the integrity of the data, making sure that they are not lost due to technical mishaps, and that the right people can access the data at the appropriate time;
- » Looks forward to the future, making it clear that you should provide detailed and structured documentation to be able to share your data with other colleagues and prepare them for long-term availability.



To make your research as time-efficient, reproducible and safe as possible, it is important that your data management is well thought through, structured, and documented.

A good data management strategy takes into account technical, organisational, structural, legal, ethical and sustainability aspects. The time invested in setting up a good data management strategy pays off when the time comes to reproduce your analysis and results.

You will be able to easily find and understand your data, increase your data's reuse potential and comply with funder mandates at the same time.

Data Management Plan

Data Management Plans (DMPs) are a key element of good data management | European Commission, 2016.

Information regarding your data management needs to be easily found and understood, not least if you are working on a project that runs over several years and involves a large team of people. In order to simplify data management, a Data Management Plan (DMP) can be created early in the research process.

A DMP is a formal document that provides a framework for how to handle the data material during and after the research project. The way a DMP will look once it is finished is not universal. It is a "living" document that changes together with the needs of a project and its participants. It is updated throughout the project to make sure that it tracks such changes over time and that it reflects the current state of your project.

A lot of diversity exists in DMPs because they are always built around the particular needs of the data collected within your project. Sometimes there are particular requirements from stakeholders that have to be answered in the DMP from stakeholders such as:

Your funders

Funders may require a Data Management Plan (sometimes called Data Publication Plan (DPP)) to get information on what data you intend to collect and whether (and how) you will make those data accessible to others. In this case you provide the funding agency with whatever information they require, to the extent that they specify. Depending on the nature of the call, such plans may include not only details on the kind and volume of data to be produced but also how the datasets will be documented and shared (along with other research outputs of the project, such as publications, program code, and educational resources). They may specify the length of the DMP or may expect you to include it in the page count of the scientific plan.

A DMP written for the funder is not always the same type of comprehensive DMP which is described in the [list of questions to this guide](#) (CESSDA, 2018a). However, the list can be used as a support when writing the DMP/DPP that the funder(s) require(s). See also the [editable version](#) (CESSDA, 2018b). Note that some funders might require that an updated DMP/DPP to be submitted as a deliverable within a specific time period. See '[Diversity in funder requirements](#)' for more information.

Your institution

Your institution may have its own policy regarding data management, including what information should be gathered and archived together with research data and publications. It's possible that your institution can support you with writing a DMP, e.g. by providing expertise or (referring to) safe storage services.

The added value of a Data Management Plan

Several researchers who I have been talking to and have looked at the Data Management Planning checklist of the Swedish National Data Service (SND) have said that doing so made them start thinking of data security, data ownership, file formats etc. before the start of their project. By doing so they avoided some possible problems that would otherwise appear later on | Ulf Jakobson, Data manager humanities, [SND](#).

A Data Management Plan (DMP) offers added value in the following ways:

Benefit 1. Useful tool to think ahead

Taking the time to plan ahead can save you a great deal of headache once the project is up and running.

Overall, a DMP helps you plan for the resources, tools, and expertise that are required to store, handle, and manage the given types and volumes of data that are expected to be collected. A DMP serves as a tool to pay careful attention to all aspects of data management. It makes you aware of possible problems at an early stage so that you can work around them. E.g. it reminds you to gain consent for future reuse and sharing from research participants.

By thinking early about various aspects of data management, you can ensure that the material is well-managed already during the data collection period. Structured and well-documented data enable others to understand the materials more easily. This, in turn, facilitates the preparation of the material for archiving, and enables further research after the end of the project.

Benefit 2. Allows for easy project management

An important function of a DMP is to work as a one-stop shop to find project-related information. Research becomes so much easier if all of your questions surrounding managing your data are being gathered in one place and project-related details are readily available rather than just vaguely remembered or simply forgotten.

A DMP is an efficient way for the researcher and his/her team to gain control over research data collection and management when the research project is up and running. Regardless of the size of the team there will be a need for easily found data-related information regarding file locations, naming conventions, standards, project description, project roles, backup regimes, versioning and so on. By writing a DMP, the researcher can ensure that the material is well-managed during the research period, which also facilitates the preparation of the material for archiving, and thus enables further research after the research project has ended. Also, it is usually easier to document research material if this is done in close proximity to the steps in the research process that create or change the material.

Project management becomes easier if you also include administrative information such as the names and [ORCIDs](#) of the Principal Investigator(s) and project members, information on which institution owns the data, registration numbers for funding and ethics board approvals. Furthermore, a lot of relevant information is kept in log books, code lists, technical reports and other documents. These documents can be referred to in the DMP together with their location information. Keeping all relevant information regarding your project in one place makes future reference a lot easier, whether that future reference is for your own thesis in three years, for an audit in five years or a historical study in fifty years.

Benefit 3. Clarifies needed budget

Data management is not free. You do not want to find yourself running out of funding before the end of the project because you have ignored or underestimated the cost of structured, detailed, and safe data management. Therefore, an important aspect of a DMP is its use in calculating how much money will be required for managing your research data during your research project.

A DMP can be useful in the process of applying for funding. Grant applications should not only include time and resources for collecting, analysing, and publishing on data in their budget, time and resources for careful documentation as well as server space, backup solutions, and documentation software need to be included as well. A DMP is also useful once funding is granted to plan and manage your expenses. Many research funders require a DMP as part of the application and decision-making process. The arguments for making data available are several, the most popular being that the data produced by public funds should be used to the greatest extent possible and available to the public. Unless there are legal, ethical or commercial barriers, data should also be openly available so that research results can be verified, replicated and reused.

Examples of Data Management cost assessments are given by the [University of Utrecht](#) (n.d.) and the Dutch Landelijk Coördinatiepunt Research Data Management ([LCRDM](#), n.d.) inspired by the '[Data management costing tool](#)' by UK Data Service, 2013.

Benefit 4. Makes data FAIRer

- » A DMP allows you to think through beforehand how to provide a dataset to a data repository which is as **FAIR** as possible. A DMP:
- » Makes structuring and documenting of your datasets simpler, thus making it easier for others as well as your future self to find and understand the material;
- » Encourages you to think about the data format which is best suited for reuse;
- » Allows you to think about the reuse license you would want to apply to your data;
- » Etc.

Benefit 5: Shows accountability

If you draw up a DMP, you are showing your affiliated institution, funders and project partners a serious approach to research data management, that includes a responsible approach towards research funds and research participants.

1.2 Research data

This expert tour guide focuses on research data management. But what is research data?

From a general perspective, research data can be described as the evidence used to inform or support research conclusions (University of Sheffield n.d.). The tangible forms this 'material' may take are e.g. "facts, observations, interviews, recordings, measurements, experiments, simulations, and software; numerical, descriptive and visual; raw, cleaned up and processed" (Van Berchum & Grootveld, 2017).

This definition combines type, form and research phase from the perspective that all manifestations of research data need to be actively managed to achieve high-quality data that have the potential to be reused. And this is exactly the perspective this tour guide adheres to.

The list below - which is based on the work of the [University of Southampton](#) (2016) - illustrates the four ways of looking at research data which are also reflected in the definition above.

Type of data

Research data can be described in many different ways. For example, they can be divided by source or by physical format. The sources of data can, for example, be registers (e.g. administrative, historical, voting results, medical, etc.), existing research data, population group(s) and communications. Physical formats of data include numerical, textual, still image, geospatial, audio, video and software. Regardless of the source and physical format of the data, data is often defined by as how they are created/captured. Examples of this includes electronic text documents, spreadsheets, laboratory notebooks, field notebooks and diaries, questionnaires, transcripts and codebooks, audiotapes and videotapes, photographs and films, examination results, specimens, samples, artefacts, slides, database schemas, database contents, models, algorithms and scripts, workflows, standard operating procedures and protocols, experimental results, metadata and other data files like e.g. literature review records and email archives.

When we speak about "new data", we mean the data that has emerged quite recently. Such data are sometimes referred to as Big Data, but both terms do not have agreed definitions.

The scholarly literature usually describes Big Data by their attributes. All of these attributes start with the letter "V" and they are Volume, Velocity and Variety (Couper, 2013).

- » Volume means that Big Data are very large and that processing them demands great computational power.
- » Velocity stands for the fact that Big Data are produced successively and new data emerge every moment.
- » Variety reminds us that Big Data are unstructured and messy and thus not ready for immediate analysis.

Some authors add two more Vs, Veracity and Value (e.g., Wamba et al, 2015):

- » Veracity tells us that Big Data must be carefully examined from the perspective of their trustworthiness. In other words, researchers should be careful about the quality of Big Data.
- » Value means that Big Data potentially generate valuable insights that are important for decision-makers, policy-makers, researchers and various organizations.

Depending on their source, the OECD defines six categories of Big Data:

A: Data stemming from the transactions of government, for example, tax and social security systems.

B: Data describing official registration or licensing requirements.

C: Commercial transactions made by individuals and organisations.

D: Internet data, deriving from search and social networking activities.

E: Tracking data, monitoring the movement of individuals or physical objects subject to movement by humans.

F: Image data, particularly aerial and satellite images but including land-based video images.

[Social media data](#) (category D) are the data from platforms like Facebook, Twitter, Instagram or YouTube. These data are created by the users of such platforms. Researchers can access these data in three main ways: 1) Direct cooperation with the companies/platforms, 2) Buying from data resellers, 3) Via APIs (one might add web scraping to the list but most platforms/companies discourage its use).

Formats

Another way to think about research data is the format in which data types (textual, numerical, multimedia, structured, software code etc.) are stored. E.g. statistical data may be stored as SPSS (*.sav) or STATA file formats, movies as *.mpg or *.avi, structured data as *.xml or in a relational MySQL database and textual files as *.docx, *.pdf or *.rtf.

Size & Complexity

The size of the files matters and so does the complexity. Managing a relatively small and simple dataset presents different challenges from managing large, complex data files.

Research phase

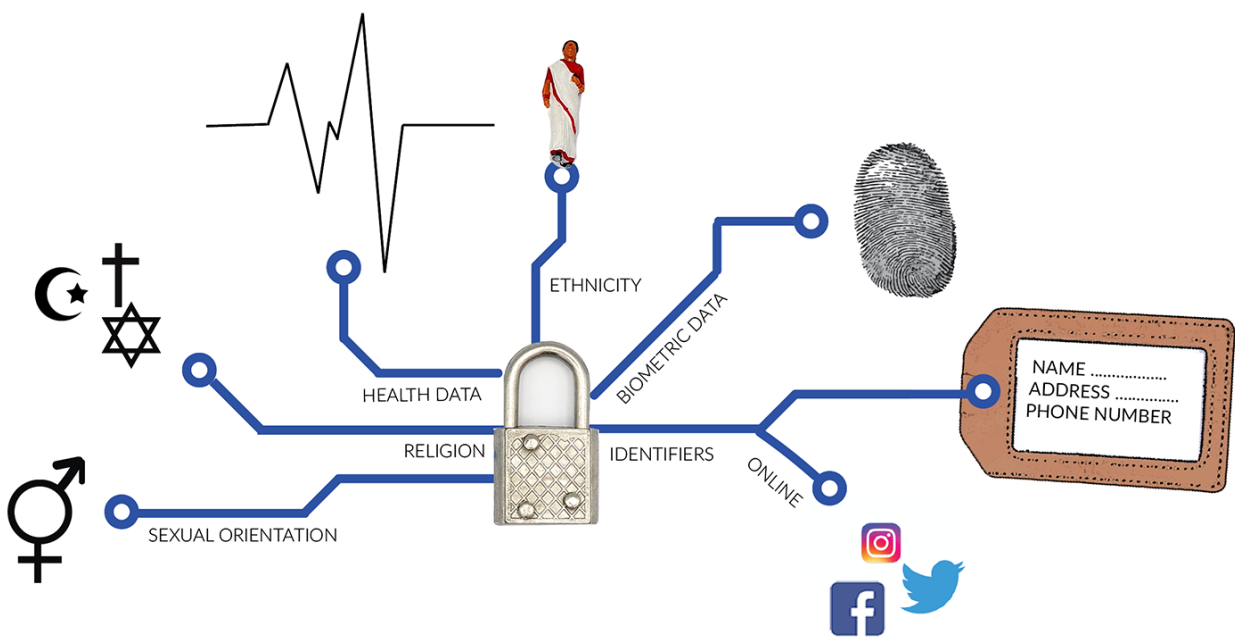
The different stages that your data travels through (raw, cleaned up, processed, analysed data) involve their own data management challenges.

1.3 Data in the social sciences

In this guide, we focus on data generated in social sciences research, both quantitative and qualitative. Notably, within the field of social sciences, you will often work with data originating from human participants. This can mean that you are handling (sensitive) personal data, which deserve special attention.

In the sections below a definition of personal data is given and our concept of quantitative and qualitative data is introduced.

Personal data



If you collect research data that enables you to identify a person, then this is classified as personal data. Within the General Data Protection Regulation ([GDPR, European Union, 2016](#)) personal data are defined as any information relating to an identified or identifiable natural person known as ‘a data subject’. It is further specified that an identifiable natural person is someone who can be identified, either directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Personal data can include a variety of information, such as names, addresses, phone numbers and IP addresses.

The GDPR applies only to the data of living persons. Data which do not count as personal data do not fall under data protection legislation, though there might still be ethical reasons for protecting this information.

Sensitive personal data

Certain personal data are considered particularly sensitive and thus require specific protection when they reveal information that may create important risks for the fundamental rights and freedoms of the involved individual. Examples of sensitive personal data include data revealing religion affiliation, sexual orientation, or racial or ethnic origin. Within the GDPR the following categories are defined as 'special categories of personal data':

- » Racial or ethnic origin;
- » Political opinions;
- » Religious or philosophical beliefs;
- » Trade union membership;
- » Genetic data;
- » Biometric data;
- » Data concerning health;
- » Data concerning a natural person's sex life or sexual orientation.

There are other data which may contain sensitive information that which does not fall under the special categories of personal data but should still be treated like as such, including, for example, confidential business data and confidential state security data.

Quantitative and qualitative data

Like with research data in general, social sciences data cover a broad range of materials, from structured numerical datasets to interviews, field notes, and documents collected for ethnographic studies, for instance. In this guide, we look at quantitative and qualitative data separately, though both can, of course, be collected during the same study.

In the table below the main attributes of both types of data are shown. Even though an attribute is described in one of the columns it does not imply that it cannot exist in the other.

Type	Quantitative data	Qualitative data
------	-------------------	------------------



General description	<p>In quantitative research, the gathered information is in numerical form. Quantitative research is used to quantify behaviour, attitudes or opinions. The goal of quantitative research is often to test ideas stated at the start of the research, to formulate facts and uncover patterns.</p>	<p>Qualitative research is primarily exploratory research. It gathers information that is not in numerical form. The goal of qualitative research is often to develop (new) ideas and a deeper understanding not achievable by numerical scores.</p>
Data attributes	<p>Data are expressed in numbers that can be assessed using statistical analyses.</p>	<p>Data are expressed in natural language, often textual or visual.</p>
Data collection methods	<p>Quantitative data collection methods include various forms of surveys – online surveys, paper surveys, mobile surveys and kiosk surveys, face-to-face interviews, telephone interviews, website interceptors, online polls, experiments and systematic observations. In most cases it generalizes results from a larger sample population.</p>	<p>Qualitative data collection methods include photography, audio recordings, video, unstructured interviews, semi-structured interviews, open-ended questionnaires, diary accounts, focus groups (group discussions), individual interviews and unstructured observations. The sample size is typically smaller than quantitative samples.</p>
Example dataset	<p>Description: Study on migrations patterns in the Summer Olympics between 1948 and 2012. The dataset covers approximately 40,000 athletes and contains information on the country they represented as well as their country of birth (open access, in English).</p> <p>Reference: Jansen, J. (Erasmus University Rotterdam) (2017): Foreign-born Olympic athletes 1948 - 2012. DANS. https://doi.org/10.17026/dans-2xf-pyqp</p>	<p>Description: Interview with a survivor of the second world war extermination camp Sobibor (open access, in English).</p> <p>Reference: Leydesdorff (copyright on the interview), prof. dr. S. (Universiteit van Amsterdam - dep. of Arts, Religion and Culture); Huffener (project manager), M. (Stichting Sobibor) (2012): Project 'Long shadow of Sobibor' Survivors: Interview 01 Thomas Blatt. DANS. https://doi.org/10.17026/dans-x8h-fwjg</p>

(Sensitive) personal data and the guide

Tips for handling (sensitive) personal data are present throughout this guide. In particular, we would like to point out the following:

- » In the [chapter on storing data](#), you will find measures to protect (sensitive) personal data from unauthorised access with strong passwords and encryption.
- » In the [chapter on protecting your data](#), you will learn how a combination of gaining consent, anonymising data, gaining clarity over who owns the copyright of your data and controlling access to data can enable the ethical and legal sharing of (sensitive) personal data.

1.4 FAIR data

The attention of researchers is increasingly directed to the phases of the research lifecycle in which data are published, shared, discovered and reused. One of the perceived ways to achieve optimal reuse is to make data FAIR (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable) (Force 11, 2014; Wilkinson, et al., 2016).

The FAIR guiding principles consist of [15 facets](#) (Dutch Techcentre for Life Sciences, 2016) which describe a continuum of increasing reusability. Importantly, data should not only be FAIR for humans but also for machines, allowing, for instance, automated search and access to data. Funders like the European Commission have drafted [Guidelines on FAIR Data Management for the H2020 programme](#) (European Commission, 2016). Good data management is one way to support the FAIR principles.



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.

Steps toward FAIRer data

In this guide, we treat the FAIR principles as guidelines to a clear higher goal: the aim is to prepare your research data for optimal (re-)use from the beginning and take appropriate measures that are most likely to be successful. To achieve FAIRness, data objects should at least have:

» **A persistent identifier (PID) for the data object as a whole**

Persistent identifiers [like DOIs](#) prevent link rot. Link rot is the process by which hyperlinks stop referring to the original source through time because they are moved or deleted. Without a PID, the data object simply will not be findable let alone reusable in the long-run (see '[Data citation](#)').

» **A sufficient set of metadata**

A sufficient and standardised set of metadata (elements which describe the data) will enhance findability, interoperability, and reusability. The quality of the descriptive information regarding the data has a profound impact on their reusability. So the more documentation of the data's context, the better. As a minimum, there should be sufficient amount of metadata to make the data findable but also understandable and reusable by other researchers (see '[Documentation and metadata](#)').

» **A clear licence**

Researchers (and computers) who find a dataset should immediately know what they are allowed to do with it. Stating clear re-use rights is like having a warm 'Welcome' on the doormat of your dataset. The motto is: 'open if possible, restricted if necessary' (see '[Data licensing](#)').

One of the ways to make sure your data will not become useless in the long-run is to choose a (trusted) data repository which has these attributes built into its infrastructure for dataset submission. It is the interest of FAIR data that researchers deposit their data, along with all the documentation needed for their understanding and re-use, in a (trusted) research data archive that has an explicit goal of data preservation and the necessary expertise to store data sustainably and maintain their usability (Van Berchum & Grootveld, 2017).

Making data FAIR is a joint responsibility of researchers and data repositories. [In a comprehensive document](#), the Swiss National Science Foundation explains (SNF, n.d.) how the responsibilities of both are distinct.

In the chapter on [archiving and publishing data](#), we will guide you in making the FAIRest choice for entrusting your data.

Expert tip



How FAIR are your data?

Want to know how FAIR your data are? Have a look at the [checklist by Jones and Grootveld](#) (2017).

1.5 European diversity



Data management requirements in Europe

There are many different local, national and international DMP templates and tools that you can use to create a DMP for your own research project. At this stage, it might be good for you to check for templates or tools that best fit your own specific situation. You can ask at your university or department whether they have a DMP template. Or maybe your research funder requires a DMP in a specific format.

In the following tables, we sum up European diversity in funder requirements on Data Management Planning and provide links to DMP templates if they are available.

[Also see the online version of this guide for an up-to-date overview of European diversity in funder requirements on Data Management Planning and links to available DMP templates.](#)

EU

Funding institution	DMP requirements	DMP template?
Horizon 2020	<p>At grant submission</p> <p>Projects that take part in the Horizon 2020 Open Research Data Pilot (default all H2020 projects) are required to produce a DMP as a deliverable within six months of the start of the project.</p> <p>The DMP should include information on:</p> <ul style="list-style-type: none"> » the handling of research data during and after the end of the project » what data will be collected, processed and/or generated » which methodology and standards will be applied » whether data will be shared/made open access and » how data will be curated and preserved (including after the end of the project). <p>During the project</p> <p>The plan must be updated whenever significant changes arise, for instance (but not only) when there are:</p> <ul style="list-style-type: none"> » new data » changes in consortium policies (e.g. new innovation potential, decision to file for a patent) » changes in consortium composition and external factors (e.g. new consortium members joining or old members leaving). <p>See 'Guidelines on FAIR data management' (European Commission, 2016).</p>	<p>Yes, via DMPOnline (Digital Curation Centre, 2017)</p>
European Science Foundation	No	No
European Research Council	<p>No</p> <p>See 'Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council under Horizon 2020' (European Research Council, 2017).</p>	<p>Template (European Research Council, n.d.)</p>

Belgium

Funding institution	DMP requirements	DMP template?
Fonds voor Wetenschappelijk Onderzoek	Yes	Yes, applicants should provide detailed information following the requirements described here .
Federal State – Belgian Federal Science Policy Office (BELSPO)	Pending	Pending
Fédération Wallonie-Bruxelles – Fonds de la Recherche scientifique (F.R.S.-FNRS)/Federation Wallonia-Brussels – Fund for Scientific Research	No	No
Fédération Wallonie-Bruxelles – Direction générale de l’Enseignement supérieur, de l’Enseignement tout au long de la vie et de la Recherche scientifique (DGESVR)/Federation Wallonia-Brussels – Directorate General of Higher Education, Lifelong Education, and Scientific Research	No	No
Vlaamse Overheid – Departement Economie, Wetenschap & Innovatie (EWI)/Government of Flanders – Department of Economy, Science, and Innovation	No	No

Croatia

Funding institution	DMP requirements	DMP template?
Croatian Science Foundation	No	No

Czech Republic

Funding institution	DMP requirements	DMP template?
Czech Science Foundation	No	No
Technology Agency of the Czech Republic	No	No
Ministry of Education Youth and Sports - R&D support programmes	No	No
Operational Programme Research, Development and Education (EU Funds, managed at the Ministry of Education, Youth and Sports)	No	No
Ministry of Culture - Applied research programmes	No	No

Finland

Funding institution	DMP requirements	DMP template?
The Academy of Finland	Highly recommended. The Academy of Finland requires that the data management plan is made using the DMP tool DMPTuuli.	Yes, via DMPTuuli (2017)
Business Finland	Highly recommended.	Yes, via DMPTuuli (2017)
The Finnish Foundation for Alcohol Studies	Highly recommended.	Recommends a DMP based on the requirements of the Finnish Academy. See ' best practices ' (Academy of Finland, 2017)
Kone Foundation	Highly recommended.	No, but see more information: Koneen Säätiö (n.d)
The Finnish Work Environment Fund	Highly recommended.	No

Germany

Funding institution	DMP requirements	DMP template?
Deutsche Forschungsgemeinschaft (DFG)	There is a general recommendation, but not a requirement. Requirements may be (and are) made on an individual program level. See DFG Guidelines on the Handling of Research Data (n.d.).	No
Bundesministerium für Bildung und Forschung (BMBF)	There are no general requirements. Requirements are made on a “per call” basis (some calls have requirements, others don’t). See Bundesministerium für Bildung und Forschung (n.d.).	No
Volkswagen Stiftung	No	No
Fritz-Thyssen-Stiftung	No	No
Hans-Böckler-Stiftung	No	No

The Netherlands

Funding institution	DMP requirements	DMP template?
NWO	Yes, both at grant submission and when the grant has been awarded.	Yes, see Netherlands Organisation for Scientific Research (2016)
KNAW	Yes, both at grant submission and when the grant has been awarded.	No, but for more information see Koninklijke Nederlandse Akademie van Wetenschappen (n.d.)
ZonMw	At grant submission information on whether data will be reused or collected needs to be provided. Also, a final DMP is required when the grant has been awarded.	Yes, see ZonMw (n.d.)

Also, you may want to have a look at the [overview of DMP templates](#) used by Dutch universities (Landelijk Coördinatiepunt Research Data Management (2017)).

North Macedonia

Funding institution	DMP requirements	DMP template?
Ministry of education and science of North Macedonia	No	No

Norway

Funding institution	DMP requirements	DMP template?
Research council of Norway	Starting in 2018, all Research Council-funded projects that generate data will as a general rule need to have a data management plan (DMP).	Yes
SkatteFUNN	Yes, during the project	No

The Research Council of Norway published their revised policy in December 2017, announcing that DMP requirements will be incorporated into calls for proposals starting in 2018. The main rule is that all Research Council-funded research projects that generate data must have a DMP. The DMP should be a living document that is updated throughout the project period, and should cover: which data are to be generated, how the data are to be described, where the data will be stored and whether and how they may be shared. The plan should be made public. The DMPs must be in place when projects submit their revised grant applications, but they will not be part of the Research Council's application review process.

The institutions are responsible for determining how they will provide access to the data. Under certain circumstances, the Research Council will require storage of the data in specific national or international archives (e.g. at NSD - Norwegian Centre for Research Data).

Serbia

Funding institution	DMP requirements	DMP template?
Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije (Ministry of Education, Science and Technological Development of the Republic of Serbia), Fond za nauku Republike Srbije (Science Fund of the Republic of Serbia)	No	No

Slovenia

Funding institution	DMP requirements	DMP template?
Slovenian Research Agency	No	No

Sweden

Funding institution	DMP requirements	DMP template?
Formas	No	No
Riksbankens Jubileumsfond (RJ)	No	No
Forte	No	No
Vetenskapsrådet (VR)	No	No
Knut and Alice Wallenberg Foundation	No	No
Marianne and Marcus Wallenberg Foundation	No	No
Marcus and Amalia Wallenberg Foundation	No	No
Stint	No	No

In Sweden, good data management is considered a key component for open access to research data, as well as a foundation for FAIR data. For the Swedish Research Council (Vetenskapsrådet, VR), a data management plan is a requirement for an approved application for funding, for all applications as of 2019. Other national public research funding bodies also require data management plans in new calls for applications. This applies to Formas, Forte, and Riksbankens Jubileumsfond, which require a DMP to be set up when funding is granted.

VR has appointed a national reference group, consisting of researchers, representatives for universities, archives, libraries, and research funding bodies, to collaborate on the work on data management plans. The national reference group has four working groups. These groups address concepts and definitions, stakeholders, DMP users, and specifications for a national tool for data management plans. Work is already in progress for a national DMP tool, and the plan is to implement a first “light” version of the tool in December 2019.

The Swedish National Data Service provides a guide and a [template for a data management plan](#).

Some universities across Sweden have, on their websites, recommendations on using SND's checklist or other organisations' DMP tools.

Switzerland

Funding institution	DMP requirements	DMP template?
The Swiss National Science Foundation	Yes, as of October (2017), data management plans are an integral part of project funding applications submitted to the SNSF. The data management plans do not need to be finalised by the submission deadline. The submitted DMP is considered a notice of intention. It is a requirement for any transfer of funding. Researchers are encouraged to adapt its contents as the project evolves. A final version must be made available at the end of the grant and will then be made available on the SNSF's P3 database. See the Guidelines for researchers (SNF, 2017).	Yes, see Data Management Plan (SNF, 2017)

UK

Funding institution	DMP requirements	DMP template?
Wellcome Trust	<p>Yes, at grant submission:</p> <p>A DMP should be provided for grant applications for projects that aim to create a database resource or will generate significant datasets that could be shared. For such an application, you need to include:</p> <ul style="list-style-type: none"> » The data outputs your research will generate » When you intend to share your data » Where your data will be made available » How your data will be accessible to others » Whether limits to data sharing are required » How key datasets will be preserved » Resources required <p>See 'Developing an outputs management plan' (Wellcome (n.d.))</p>	Yes, via DMP Online (Digital Curation Centre, 2017)
Economic and Social Research Council	Yes, at grant submission. See the ESRC Research Data Policy (ESRC, n.d.)	Yes, via DMP Online (Digital Curation Centre, 2017)
Cancer Research UK	Yes, at grant submission	Yes (Cancer Research (n.d.))
Department for International Development	Yes, at grant submission	Yes (Department of International Development (2013))

Open Data and Open Science policies in Europe

For a snapshot of various Open Data and Open Science policies, as they currently stand throughout Europe, you can have a look at [this living report](#) (SPARCEurope & Digital Curation Centre, 2017).

1.6 Adapt your DMP: Part 1



The Data Management Plan (DMP) is an important tool to structure the research data management of your project. After working on each chapter you should be able to answer part of the questions which make up a DMP.

This is the first of seven 'Adapt your DMP' sections in this tour guide. When you have finished the chapter on data management planning, you can start filling in the 'Overview of your research project' section. Below you can see what elements and corresponding questions are generally included in that section. You can select appropriate questions and answer them to adapt your own DMP.

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Alternatively, an [online DMP solution](#) developed by NSD is available. To access login with EduGain or google account is necessary. Currently, two templates (H2020 and a general one) are offered.

DMP Questions

Title of the project / study

What is the title of your project? Give a short description.

Date and version of this plan

- » What is the date of this DMP version?
- » How do you discern between versions of your DMP?

Description of the project

- » What is the nature of the project?
- » What is the research question?
- » What is the project timeline?

Origin of the data

- » What kind of data will be used during the project?
- » If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated?
- » If you are collecting new data can you clarify why this is necessary?

Principal and collaborating researchers

Principal researchers

- » Who are the main researchers involved?
- » What are their contact details?

Collaborating researchers (if applicable)

- » What are their contact details and their roles in the project?

Funder (if applicable)

If funding is granted, what is the reference number of the funding granted?

Data producer

Which organisation has the administrative responsibility for the data?

Project data contact

Who can be contacted about the project after it has finished?

Data owner(s)

- » Which organisation(s) own(s) the data?
- » If several organisations are involved, which organisation owns what data?

Roles

- » Who is responsible for updating the DMP and making sure that it's followed?
- » Do project participants have any specific roles?
- » What is the project time line?

Costs

- » Are there costs you need to consider to buy specific software or hardware?
- » Are there costs you need to consider for storage and backup?
- » Are potential expenses for (preparing the data for) archiving covered?

Examples of DMP questions and answers

For inspiration of filled in DMPs look at some example DMPs we have prepared. Both DMPs are based on a fictional research project with a basis in reality. For each topic of the DMP, there are example questions and answers where applicable. The examples are not country specific. Some of the information is generic.



Qualitative data

During this project, in-depth interviews with teachers in primary school will be held. The project has just started.

[Download examples](#)



Quantitative data

The project concerns a survey which is conducted in order to identify how the evolution of society affects attitudes and behaviour. The project is still running.

[Download examples](#)

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 2

Organise & Document

Contents

Main take-aways	33
2.1 Designing a data file structure	34
2.1.1 Organisation of variables	39
2.2 File naming and folder structure	43
2.3 Documentation and metadata	47
2.4 Adapt your DMP: part 2	57
Sources and further reading	58

[View the online version of this chapter](#)

Main authors of this chapter

Jindrich Krejčí, Czech Social Science Data Archive (CSDA)

Johana Chylikova, Czech Social Science Data Archive (CSDA)

Katja Fält, Finnish Social Science Data Archive (FSD)

Introduction



In this chapter¹, we provide you with tips and tricks on how to properly organise and document your data and metadata. We begin by discussing good practices in designing an appropriate data file structure, file naming and organising your data within suitable folder structures. You will see how the way you organise your data facilitates orientation in the data file, contributes to the understanding of the information contained and helps to prevent errors and misinterpretations.

In addition, we will focus on what counts as an appropriate documentation of your data. Development of rich metadata is required by [FAIR data principles](#) and other current standards promoting data sharing.

Main take-aways

After completing your journey through this chapter on organising and documenting your data you should:

- » Be aware of the elements which are important in setting up an appropriate structure for organising your data for intended research work and data sharing;
- » Have an overview of the best practices in file naming and organising your data files in a well-structured and unambiguous folder structure;
- » Understand how comprehensive data documentation and metadata increases the chance your data are correctly understood and discovered;
- » Be aware of common metadata standards and their value;
- » Be able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt them to your own DMP.

¹ The content of this chapter was inspired by research data management manuals, guidelines, online courses and methodological texts published by several data organisations and experts, in particular the [information provided by the UK Data Service](#) (2017a), the [“Guide to Social Science Data Preparation and Archiving”](#) by the US-based data organisation ICPSR (2012), the online course [Research Data MANTRA](#) (EDINA and Data Library, University of Edinburgh, 2017), A guide into research data management by Corti, Van den Eynden, Bishop and Woollard (2014), Krejčí’s “Introduction to the Management of Social Survey Data” (Krejčí, 2014), Gibbs (2007) and [Data Management Guidelines](#) produced and published by the Finnish Social Science Data Archive (FSD-Finnish Social Science data Archive, 2017).

2.1 Designing a data file structure

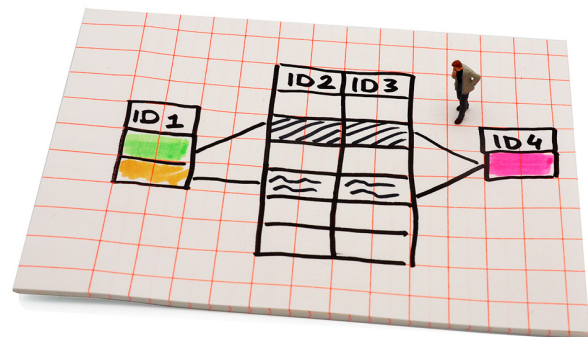


In an early stage of your research, you are faced with the question of what form your data files should take. Your initial decisions about the structure of your data files should be considered thoroughly.

Data file structure has a huge impact on the possible ways your files can be processed and analysed and once your structure has been filled with data, any changes to it are usually laborious and time-consuming.

File structure choice

Data files may have different internal structures and a research study may encompass several different data files in different relations to one another. The structure of the data file is also determined by the formatting of its content (e.g., types and organisation of variables). It provides information on relationships among different elements and parts of its content. An important part of the metadata is often embedded into the data file (e.g., in the form of variable names and variable and value labels, different kinds of notes and content of supplementary variables). So, the structure of your data also contributes to the clarity of your data documentation.



File structure choice often depends on the requirements of the software you are using and intended analysis. At the same time, your decisions about structure may define the possibilities of future data processing, choice of software and ways of data analysis.

When deciding on a data file structure, consider the following:

- » Units of analysis, possible analytical objectives and methods of analysis to be used;
- » Relations
 - » between different content items and parts of your data file;
 - » to sources of your data;
 - » to any other relevant external data and information and their structure.
- » Possibilities of building connections to other existing or future data files (future additions of new data or creation of cumulative data files);
- » Possible strategies for version control (see [‘Data authenticity and version control’](#));
- » Possible technical limitations, e.g. operability in relation to the size of the data file (consider that large and complicated structures may put high demands on both data management and computing capacities. Some software programs also have limitations with respect to the number of variables and cases they can manage);
- » The software you are going to use (this should be done also with respect to flexibility because of possible secondary analysis of your data in other software).

Designing qualitative data files



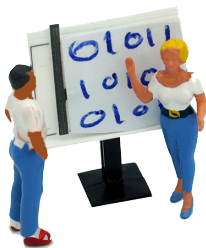
Qualitative data files emerge from many different types of research material. Such data files are texts (transcribed interviews or focus group sessions, various types of written texts, such as newspaper and magazine material, diaries etc.) or photographs, audio files (recordings of speech) or video files. Unlike quantitative data, qualitative data are not presented in the form of variables, numbers, data matrices etc. However, they must also be organised and stored in an exact manner so they are easily managed and available for use.

Usually, individual data collection events will be structured into individual files, e.g. one interview transcript, one image, or one audio recording each time make a single file. These single files are then organised into folders of similar files. Sometimes, qualitative information may also be organised into matrix structures, e.g. textual extracts from newspaper articles or diaries may be placed into a rectangular matrix, whereby further metadata and coding can be added alongside each entry.

Designing a qualitative data structure comes down to:

- » Thinking of ways to categorise data (see '[Qualitative coding](#)');
- » Developing a file naming strategy (see '[File naming and folder structure](#)');
- » Designing a comprehensive folder structure (see "[File naming and folder structure](#)").

Designing quantitative data files



In quantitative research, the content of the data often results from numerical coding in standardised questionnaires (see '[Quantitative coding](#)'). In addition, full-text answers or textual codes can be recorded into specific types of variables in quantitative data files. Quantitative researchers may also store other material, i.e. administrative data, data from social media or various texts. In this chapter, however, when we speak about quantitative data, we usually mean survey data.

Below you will find a description of three types of file structures - flat, hierarchical and relational - which are commonplace in quantitative social science. Also, two examples which clarify the concepts are presented.

Flat file

Flat (rectangular) data files are organised in long rows, variable by variable. One row is dedicated to one subject of observation and/or analysis. An ID number usually comes first. If variable values are organised column by column, we obtain a rectangular matrix.

SPSS and STATA and similar software are often used for analysing flat files. Here the structure consists of one rectangular matrix with data, accompanied by variable and value labels. In this case, each record includes the same amount of information and has the same length as all other records in the data file. If the variable is not applicable for a particular observation, it is filled with blank spaces or missing values.

Hierarchical file

Hierarchical files consist of higher-order and lower-order records which are arranged in a hierarchical structure, i.e. several lower-order units may be linked to one higher-order unit and are contained in the same data file.

If there are different levels of units in your database the flat data file can be impractical because it may include a large number of blanks and put great demands regarding the size of the file. In addition, it may also reduce the operability and clarity with regard to differentiation of types of units of analysis. Database applications like e.g. D-base, MS Access or SQL, allow structuring your data in a hierarchical order.

Relational database

The relational database is a system of several data matrices and defined associations between them.

Different other database applications, e.g., D-base, MS Access or SQL databases, allow the structuring of your data in a hierarchical order. You may also split your data into several interrelated flat files, i.e. structure your data into a relational database, and retain the ability to use statistical software mentioned above.

Example 1: Structuring a database from a household survey

If you consider a database from a household survey, there are at least two different types of units of analysis, households and individual household members. However, you may structure such database in all three ways as follows:

Hierarchical structure

If you decide on a hierarchical structure, data on the household are recorded at one level and data on household members at another level.

You can download an example of a hierarchical file in *.sav [here](#).

Relational database

Another solution is to create a relational database. Information about household members is 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 or similar. 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.

Flat file

However there could be a situation where you may need to use complete household survey data in your analysis and your software requires a flat file. In this case, you can add a household ID variable and copy particular household data for each individual member of this household. 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.

You can download an example of such a flat file in *.sav [here](#).

Example 2: SHARE - A complicated database of micro data on health, socio-economic status and social and family networks

The Survey of Health, Ageing, and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks. Surveys are organised bi-annually since 2004. SHARE currently covers 27 European countries and Israel. The SHARE database is easily accessible to the entire research community; data from the SHARE Waves 1 to 6 have been available since 2017.

Description

SHARE is targeted to individuals aged 50 or older and their households. The resulting database is quite complicated due to the following:

- » SHARE is an international survey and its data come from different countries;
- » SHARE is a panel survey repeating interviews with the same sample of households every two years, so the data come from different waves of the survey using questionnaires including both, repeated and new questions;
- » A systematic process of refreshment is implemented and new households are added to the panel at each wave. That is the reason there are two types of questionnaires: the baseline questionnaire for respondents who participate in a SHARE interview for the first time and the longitudinal questionnaire for respondents who participated in SHARE before;
- » There are different components of the survey with different sources of information collected under different data collection modes. The data collection modes include face-to-face interviews based on computer-assisted personal interviewing (CAPI) on household level, CAPI on individual level for different types of household members, paper and pencil (PAPI) drop-off questionnaires, so-called vignettes, i.e. questionnaires on respondents reactions to specific situations, physical tests, collection of dried blood spots (only in some countries), specific end-of-life questionnaires, interviewer observations and generated variables;
- » Different types of respondents answer different parts of an interview for household: (1) family respondent, (2) financial respondent, (3) household respondent; in addition, e.g., in case of physical or cognitive limitations of dedicated respondent, it is possible to organise a 'proxy interview' with another member of the household (proxy respondent);
- » The survey and its database are also structured by topics.

Unique identifiers

The features of the surveys which are described above, result in a complicated relational database.

Up to 30 different data modules are available for each wave of the survey, for each of participating countries and for databases combined across time and countries. Moreover, there are also two levels of data based on units of observation, households or individuals. Data from different modules and/or waves may be merged using the following set of unique identifiers:

- » For merging data on an individual level the variable "mergeid" provides a unique and non-changing person identifier for all waves. It has the format "CC-hhhhhh-rr" (e.g. "AT-070759-01"), where CC refers to the short country code (here: "AT" for Austria), "hhhhhh" are digits to identify the household, and "rr" is the respondent identifier within each household.
- » For merging data on the household level there is a set of variables hhid`w, where `w indicates the respective wave. hhid`w has the following format "CC-hhhhhh-S" (e.g. "AT-070759-A"), where "CC" refers to the short country code, "hhhhhh" is the household identifier, and "S" identifies possible split households, i.e. the household of a panel member who moved out of a previous household. In case of a household split there is not only an "A"-suffix but also "B", "C", etc.

In addition, there are several 'Special Data Sets', e.g. interviewer survey, country-specific projects to link SHARE data with selected administrative records and 'Biomarkers' (objective health measures or a retrospective panel about working life histories of SHARELIFE respondents (SHARE Wave 3)).

For purposes of analysis, SHARE provides a very extensive set of weights (See Weighting). Which weights to use really depends on the concrete research question, i.e. the cross-sectional or longitudinal nature of the study, the waves under investigation, the unit of analysis (household or individual), and the reason for weighting sample observations (SHARE, 2017: 34).

easySHARE for training purposes

Working with the SHARE panel data is very demanding. Thus, in addition to the standard SHARE database, also a longitudinal data set "easySHARE" has been created for training purposes. It contains only selected variables merged into a single data file, making it more user-friendly. However, for deeper analysis, a standard database is necessary.

Dive in deeper?

We have a subtopic prepared for you on [organising variables](#). Here you will find tips on how to build the internal structure of quantitative data files by organising, naming and labeling variables.

Alternatively, you can proceed to the section on designing [file names and folder structures](#).

2.1.1 Organisation of variables

Data file structure is supported by the organisation of variables. Variable names and labels contribute to the structuring of the data file, allowing to integrate part of the documentation into the data file and helping researchers to orient themselves in the structure of the data sets. At the same time, variable names should be short and should respect the usual requirements of standard software, because they are used as calling codes in software operations.

The position of variables in the data file, their names and labels should reflect the following:

» **Relations between variables**

E.g., sets of variables related to the same phenomenon (these should be placed together in a dataset, e.g. the age of all children in a household), original or derived variables (derived variables are created from other variables, e.g. the age in years is re-coded into broader categories)

» **Links to elements of the study and sources of the data**

E.g., different measurement instruments, different parts of the questionnaire, different source databases, different methods of observation, etc.

» **Types of variables**


E.g., identification variables and other supplementary variables with different specific roles, socio-demographic indicators, generated variables obtained by transformation of original information, etc.

Organising your data

Data files also include supplementary variables which facilitate orientation and management, ensure integrity, or are necessary to perform some analyses. As a rule, you should include a unique identifier (or set of identifiers) for cases (individual respondents) in the file. A unique identifier is an identification code for the case. These are usually numbers, for example, 0001, 0002, 0003 etc. To facilitate orientation, they are usually placed at the very beginning of the file.

List of variables:

SEX	Sex of respondent
AGE	Age of respondent
MARITAL	Marital status of respondent
COHAB	Do you live together with a partner?
EDUCYRS	Education I – years (of fulltime) schooling



Other variables may help to 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 waves of research, etc., especially if groups of cases distinguished by them are to be analysed in different ways.

For each variable in the data file, you should set the variable width, i.e. the number of characters or the length of the integer and fractional parts of a number. The set number of characters or digits for each variable is reserved for every case, even if they are left blank.

Naming variables

In the boxes below, basic rules for variable naming are given and an example is presented.

Basic rules for variable naming

The basic rules for variable naming are following:

- » Start with a letter. Do not start with a number, question or exclamation marks or a special character such as #, &, \$, @ (they are often reserved for specific purposes in software applications);
- » Variable names cannot contain spaces;
- » Variable names are also used as calling codes in software operations. For this reason, variables should be short and respect the usual requirements of a standard software. The standard is to not make variable names any longer than eight characters;
- » Do not use diacritics (marks above or below a letter) or national specific characters;
- » Make them meaningful (so they can be used for better orientation in the data files).

There are three basic approaches to naming variables:

- » Using numeric codes that reflect the variable's position in a system (e.g. V001, V002, V003...);
- » Using codes that refer to the research instrument (e.g. question number in a questionnaire: Q1a, Q1b, Q2, Q3a...);
- » Using mnemonic names that refer to the content of variables (e.g. BIRTH for the year of birth, AGE for respondent's age etc.). The word mnemonic means "memory aid".

Variable labels

Variable labels provide a short description of the variable name. These can be longer than the recommended eight characters for variable names. Although size limits are less strict here, it is advisable to keep variable labels rather brief and find an adequate compromise between clarity and the size of the label. Keep in mind that many analytical outputs are provided in tables. Thus, excessively lengthy labels can result in large and impractical tabulations. The size of the labels may also complicate format conversions. In some analytical outputs or after conversions, only a part of a lengthy label is kept. The loss of the remainder of the variable label may make the label incomprehensible.

Examples of variable labels include a short or full version of the question, or a question code if variable names are not constructed around them. E.g.:

- » The variable label is adapted from the number and question-wording from the questionnaire: "B10 - How old are you?";
- » The descriptive label is "Age of a respondent";
- » Schematically this becomes: "Respondent: AGE".

To reach the widest audience possible, the preferred language for variable naming is English.

Labels for variable values

Variables have two or more values (a variable with only one value is called a constant and in fact, is not a variable). Sometimes you must assign labels to values of variables. You do not need to assign labels to values of continuous variables like age (in years), height (in metres) or weight (in kilograms), because their units are generally known. This is different for nominal and ordinal variables. A nominal variable like gender has two values, usually represented by 0 and 1 in data. You should assign labels “male”/“female” to these two values, so you and another researcher who might use the data would know which value represents which gender. The same applies to ordinal scales, for example, agree-disagree scale with values 1, 2, 3, 4 and 5, where 1 represents “completely disagree” and 5 “completely agree”. You must label these values so you and others know what degree of dis/agreement the numbers represent.

Example

Two different concepts of variable naming and labelling in the data file from the International Social Survey Programme

The International Social Survey Programme (ISSP) is a continuing, long-term international programme of survey research on important sociological topics. It brings together pre-existing, social science projects and coordinates research goals, thereby adding a cross-national perspective to the individual, national studies. Established in 1984, it now has almost 50 member countries. The ISSP surveys are organised annually.

Each ISSP survey contains two international modules:

- » **ISSP thematic module**

A specific topic of the survey is selected for each year. There are about ten topics, which are repeated at regular intervals. However, sometimes a topic is skipped or replaced by a new one.

- » **ISSP background variables module**

These include a set of harmonised sociodemographic variables. This module is repeated every year. However, there are also frequent changes in this set of variables.

Two different concepts of variable naming and labelling are used for these two modules.

Table: Excerpt from the variable list of the international dataset from ISSP 2009 on 'Social Inequalities' ([ISSP Research Group, 2017](#)).

Variable name	Variable label
<i>ISSP 2009 thematic module variables</i>	
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 every day
V77	Q25b Describe yourself as <14-15-16> years old: I performed to the best of my ability
<i>ISSP background variables</i>	
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

In the table we see two approaches to variable labelling:

» **Simple variable names**

The first thematic part of the file contains simple variable names (numeric codes). The information on the numbers of the questions in the common international questionnaire is included in variable labels. It supports better user orientation in the data file. The question numbers are followed by a literal question, sometimes shortened adequately to remain comprehensible and follow the rule of keeping 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, the currency used), general names come in inequality signs.

» **Mnemonic names of variables**

The second part contains background variables and uses mnemonic names of variables referring to their contents. These background variables are not directly linked to the wording of questions in the international questionnaire but are instead constructed from national versions of data. Their names refer 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). Moreover, the set of mnemonic names of background variables is standardised across different ISSP surveys, which allows easier merging of ISSP data files across time and construction of time-series databases.

TIP! Mnemonic variable names may help to establish links between sets of variables within a data file. In addition, in repeated surveys, if the same naming convention of mnemonic names is used, it makes easier merging data over time.

2.2 File naming and folder structure

To enable you to identify, locate and use your research data files efficiently and effectively you need to think about naming your files consistently and structuring your data files in a well-structured and unambiguous folder structure.

File naming strategy

Two important starting points for your file naming strategy are:

» **A file name is a principal identifier of a file**

Good file names provide useful clues to the content, status and version of a file, uniquely identify a file and help in classifying and sorting files. File names that reflect the file content also facilitate searching and discovering files. In collaborative research, it is essential to keep track of changes and edits to files via the file name.

» **File naming strategy should be consistent in time and among different people**

In both quantitative and qualitative research file naming should be systematic and consistent across all files in the study. A group of cooperating researchers should follow the same file naming strategy and file names should be independent of the location of the file on a computer.

Below, best practice and examples of useful file names are given.

Elements in a file name

Common elements that should be considered (UK Data Service, 2017b) when developing a file naming strategy:

- » Version number (also see ['Data authenticity'](#));
- » Date of creation (date format should be YYYY-MM-DD);
- » Name of creator;
- » Description of content;
- » Name of research team/department associated with the data;
- » Publication date;
- » Project number.

Best practice

According to the UK Data Archive ([UK Data Service, 2017b](#)), a best practice in naming files is to:

- » Create meaningful but brief names;
- » Use file names to classify types of files;
- » Avoid using spaces, dots and special characters (& or ? or !);
- » Use hyphens (-) or underscores (_) to separate elements in a file name;
- » Avoid very long file names;
- » Reserve the 3-letter file extension for application-specific codes of file format (e.g. .doc, .xls, .mov, .tif);
- » Include versioning of file names where appropriate.

File naming for qualitative data

Several aspects of naming that are particularly important for qualitative data ([Finnish Social Science Data Archive, 2016](#)):

- » If you have large numbers of files of different types, you should produce a document describing the file naming convention used for the research;
- » Background information about each item (individual interview, focus group, photograph etc.) is usually indicated in the file names. Nevertheless, you should always present background information in separate documents.

Consistency of naming

The benefit of consistent naming of data files is that it is easier to identify all files connected to one data collection event (e.g. one interview). The files related to one collection event (e.g. audio tape, its transcription and photographs that were taken by the interviewee) can be connected by the file name.

The most convenient way is to give all files connected to the same event an 'event identifier' in the beginning of the name, that is, in the first part of the name. The latter part of the name can be used to convey the specifics, for instance, whether it is an audio tape, transcription or a still image:

Example:

- » 20130311_interview2_audio.wav
- » 20130311_interview2_trans.rtf
- » 20130311_interview2_image.jpg

Documenting data file conventions

An example of how to document the data file conventions you use:

<date><type><ID1><gender><age><municipality><datatype><ID2>

where:

- » <date> is the date on which the data were collected (date format should be YYYY-MM-DD);
- » <type> specifies the type of event/data material;
- » <ID1> is the ID of the collection event;
- » <gender> is the gender of the interviewee;
- » <age> is the age of the interviewee;
- » <municipality> is the municipality of residence of the interviewee;
- » <datatype> specifies the type of data the file contains, for instance, "trans" means transcription, "audio" means audio recording, and "image" means photograph;
- » <ID2> is the ID number used to separate the images connected to the collection event.

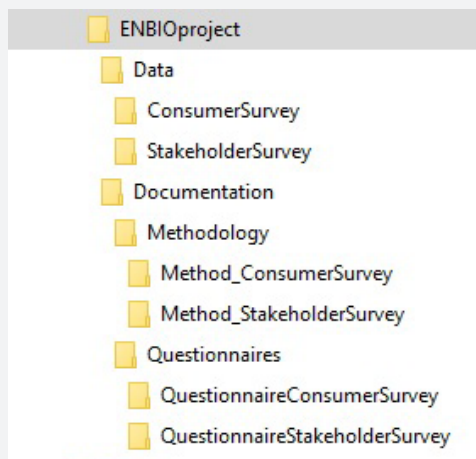
Folder structure

Structuring your data files in folders is important for making it easier to locate and organise files and versions. A proper folder structure is especially needed when collaborating with others.

The decision on how to organise your data files depends on the plan and organisation of the study. All material relevant to the data should be entered into the data folders, including detailed information on the data collection and data processing procedures.

Consider the best hierarchy of your files and decide whether a deep or shallow hierarchy is preferable. If you have several independent data collections, it is advisable to create a separate data folder for each collection. For inspiration, have a look at the examples below.

Survey data



For this survey, data and documentation files are held in separate folders. Data files are further organised according to data type and then according to research activity.

Documentation files are organised also according to the type of documentation file and research activity. It helps to restrict the level of folders to three or four deep and not to have more than ten items on each list.

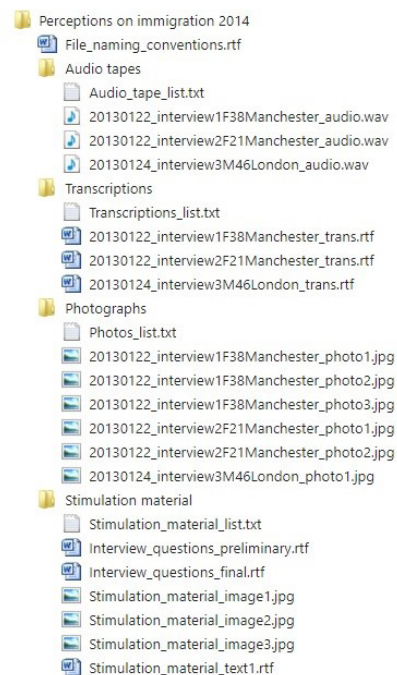
Qualitative data files

In this example, the data contain audiotapes of the interviews, interview transcripts, stimulation material shown to the research subjects, and photographs taken by the subjects.

Data files are files connected to the same interview event conducted on the 22nd of January 2013. The latter part of the name reveals the specifics of the file.

In this case, “audio” means audio tape and “trans” a transcription of the audio tape. However, background information must never be stored in the file name only.

This example was taken from UK Data Service (2017b).



TIP: Batch renaming of automatically generated files



Batch renaming is organising research data files and folders in a consistent and automated way with software tools (also known as mass file renaming, bulk renaming).

Batch renaming software exists for most operating systems. See the box below for examples.

Batch renaming tools

Examples of batch/bulk renaming tools include:

Windows:

- » [Ant Renamer](#);
- » [RenameIT](#);
- » [Bulk Rename Utility](#).

Mac:

- » [Renamer 5](#);
- » [Name Changer](#).

Linux:

- » [GNOME Commander](#);
- » [GPRename](#).

Unix:

- » The use of the `grep` command to search for regular expressions.

It may be useful to rename files in a batch when:

- » Images from digital cameras are automatically assigned base filenames consisting of sequential numbers;
- » Proprietary software or instrumentation generate crude, default or multiple filenames;
- » Files are transferred from a system that supports spaces and/or non-English characters in filenames to one that doesn't (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

How to ... use Bulk Rename Utility

Follow the steps [in this video](#) to use Bulk Rename Utility to batch rename your files.

2.3 Documentation and metadata



I have never documented my data before. I have both qualitative and quantitative data and I work on a collaborative project. Where do I start?

How to start?

1. Do not panic. Much documentation is simply good research practice, so you are probably already doing much of it.
2. Start early! Careful planning of your documentation at the beginning of your project helps you save time and effort. Do not leave the documentation for the very end of your project. Remember to include procedures for documentation in your data management planning.
3. Think about the information that is needed in order to understand the data. What will other researchers and re-users need in order to understand your data?
4. Create a separate documentation file for the data that includes the basic information about the data. You can also create similar files for each data set. Remember to organise your files so that there is a connection between the documentation file and the data sets.
5. Plan where to deposit the data after the completion of the project. The repository probably follows a specific metadata standard that you can adopt.
6. Document consistently throughout the project. Data documentation gives contextual information about your dataset(s). It specifies the aims and objectives of the original project and harbours explanatory material including the data source, data collection methodology and process, dataset structure and technical information. Rich and structured information helps you to identify a dataset and make choices about its content and usability.

TIP: Use English for documentation. It increases the chance your data are understood and reused.

Systematically documented research data is the key to making the data publishable, discoverable, citable and reusable. Clear and detailed documentation improve the overall data quality. It is vital to document both the study for which the data has been collected and the data itself. These two levels of documentation are called project-level and data-level documentation.

Project-level documentation

The project-level documentation explains the aims of the study, what the research questions/ hypotheses are, what methodologies were being used, what instruments and measures were being used, etc. In the following boxes, the questions that your project-level documentation should answer are stated in more detail.

1. For what purpose was the data created

Describe the project history, its aims, objectives, concepts and hypotheses, including:

- » The title of the project;
- » Subtitle;
- » Author(s)/creator(s) of the dataset;
- » Other co-workers and their roles (person, research group or organization that participated in the study and their roles);
- » The institution of the author(s)/creator(s);
- » Funders;
- » Grant numbers;
- » References to related projects;
- » Publications from the data.

2. What does the dataset contain?

Describe what is in a dataset:

- » Kind of data (interviews, images, questionnaires, etc.);
- » File size (in bytes), file format of the data files and relationships between files;
- » Description of data file(s): version and edition, structure of the database, associations, links between files, external links, formats, compatibility.

3. How was the data collected?

Describe how the data was acquired:

- » The methodology and technique used in collecting and creating the data;
- » Description of all the sources the data originate from (What is the subject of study? E.g. periodicals, datasets created by others?) together with an explanation of how and why it got to the present place (provenance);
- » The methods/modes of data collection (for example):
 - » The instruments, hardware and software used to collect the data;
 - » Digitisation or transcription methods;
 - » Data collection protocols;
 - » Sampling design and procedure;
 - » Target population, units of observation.

4. Who collected the data and when?

Describe the:

- » Data collector(s);
- » Date of data collection;
- » Geographical coverage of the data (e.g. Nation).

5. How was the data processed?

Describe your workflow and specific tools, instruments, procedures, hardware/software or protocols you might have used to process the data, like:

- » Data editing, data cleaning;
- » Coding and classification of data.

6. What possible manipulations were done to the data?

Describe if and how the data were manipulated or modified:

- » Modifications made to data over time since their original creation and identification of different versions of datasets;
- » Other possible changes made to the data;
- » Anonymisation;
- » For time series or longitudinal surveys: changes made to methodology, variable content, question text, variable labelling, measurements or sampling.

7. What were the quality assurance procedures?

Describe how the quality of the data has been assured:

- » Checking for equipment and transcription errors;
- » Quality control of materials;
- » Data integrity checks;
- » Calibration procedures;
- » Data capture resolution and repetitions;
- » Other procedures related to data quality such as weighting, calibration, reasons for missing values, checks and corrections of transcripts, transformations.

8. How can the data be accessed?

Describe the use and access conditions of the data:

- » Where the data can be found (which data repository);
- » Permanent identifiers;
- » Access conditions such as embargo;
- » Parts of the data that are restricted or protected;
- » Licences;
- » Data confidentiality;
- » Copyright and ownership issues;
- » Citation information.

Data-level documentation

Data-level or object-level documentation provides information at the level of individual objects such as pictures or interview transcripts or variables in a database. You can embed data-level information in data files. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file. And for quantitative data variable and value names can be embedded within the data file itself.

Quantitative data

Variable-level annotation should be embedded within a data file itself. If you need to compile an extensive variable level documentation, you can create it by using a structured metadata format.

Data-level documentation for quantitative data

For quantitative data document the following information is needed:

- » Information about the data file
Data type, file type, and format, size, data processing scripts.
- » 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. Variable labels should:
 - » Be brief with a maximum of 80 characters;
 - » Indicate the unit of measurement, where applicable;
 - » Reference the question number of a survey or questionnaire, where applicable.

Example of a variable and variable label

Variable: 'Q11eximp'

Variable label: 'Q11: How important is exercise for you?'

Value labels: 1: Very unimportant. 2. Unimportant. 3. Neutral. 4. Important. 5. Very important.

The label gives the unit of measurement and a reference to the question number (Q11).

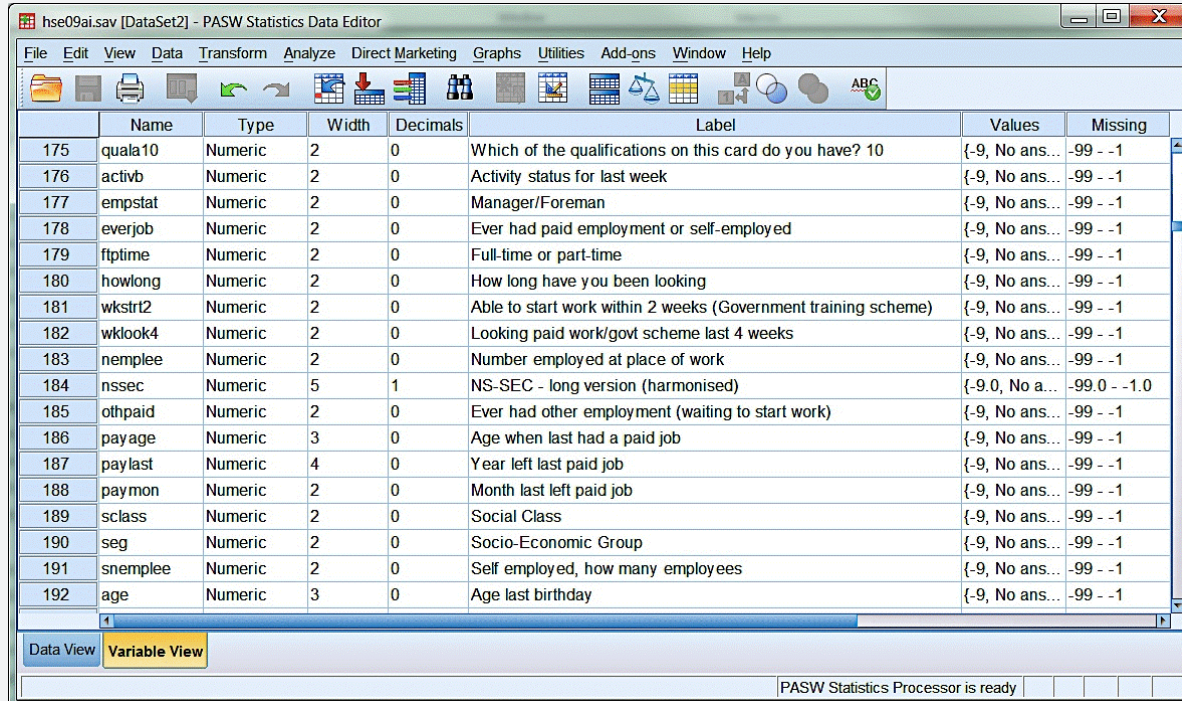
- » Information about the cases in the file
- » A specification of each case (units of research like e.g. a respondent) if applicable.
- » Names, labels and descriptions for variables, records and their values
- » Description of the missing values at each variable
- » Description of the weighting variable
- » Explanation or definition of codes and classification schemes used

Storing documentation

Whenever possible, embed data documentation within a file. See the following example.

Example of embedded data documentation

In this example from the UK Data Service (2017c), you see two SPSS tabs: Data view and Variable view, the tab which is visible right now.



	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 - -1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 - -1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 - -1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 - -1
182	wklook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 - -1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 - -1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9,0, No a...	-99,0 - -1,0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 - -1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 - -1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 - -1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 - -1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 - -1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 - -1
191	sneemplee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 - -1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 - -1

Qualitative data

Background and contextual information and participant details of interviews, observations or diaries can be described at the beginning of a file as a header or summary page.

Data-level documentation for qualitative data

For qualitative data document the following information is needed:

- » Textual data file (for example, interview)
- » Key information of participants such as age, gender, occupation, location, relevant contextual information);
- » For qualitative data collections (for example image or interview collections) you may wish to provide a data list that provides information that enables the identifying and locating of relevant items within a data collection:
 - » The list contains key biographical characteristics and thematic features of participants such as age, gender, occupation or location, and identifying details of the data items;
 - » For image collections, the list holds key features for each item;
 - » The list is created from an initial list of interviews, field notes or other materials provided by the data depositor.

Example of data level documentation of textual data

For textual data, background data are systematically entered at the beginning of each data unit (e.g. interview transcript) in a standardised manner.

The following example from the [Finnish Social Science Data Archive](#) presents a typical transcript of an interview with only one interviewee. The transcript of each interview in the data has been saved in a separate file, often in .rtf or .doc(x). Background data fields are entered in the following manner at the beginning of each transcription file.

Beginning of the transcript file

Interview date: 08.02.2013 [=8 February 2013]

Interviewer: Matt Miller

Pseudonym of interviewee: Ian (not the real first name of the interviewee)

Occupation of interviewee: Journalist

Age of interviewee: 32

Gender of interviewee: Male

Audiovisual data files

For some types of data (image, audio or video files) the file format does not always allow recording background information in the beginning of the data file. In such cases, the best practice is to store background information in a manually created data list or a separate text file: a data list which accompanies the data collection.

- » Provide the following information on each image: creator, date, location, subject, content, copyright, keywords, equipment used;
- » Some image files have embedded technical metadata (You may use tools to extract technical metadata from images, such as [ExtractMetadata.com](#) (n.d.)).

Example of a data list

In this case - [shown on the site of the Finnish Social Science Data Archive](#) (2016) - the background data fields are manually entered in table form using Excel (or Open Office Calc program). The data collected were video-recorded interviews. The data list contains background information related to the interviewee and the interview event as well as information on the model and brand of the camera used and the length of the video (in minutes).

See also another [data list example](#) from the UK Data Service (2017c)

1	Interview videos 2012								
2	File name	Interview date	Interviewer	Interviewee's name	age	gender	occupation	Camera used for the video	Duration of the video
3	Peter_1.avi	12.4.2012	Matt Miller	Peter Herald	37	Male	Barkeeper	Panasonic HC-V10	2:45
4	Peter_2.avi	12.4.2012	Matt Miller	Peter Herald	37	Male	Barkeeper	Panasonic HC-V10	5:05
5	Lisa_1.avi	17.4.2012	Matt Miller	Lisa Smith	43	Female	Author	Canon XF305	10:12
6	Mary_1.avi	22.4.2012	Matt Miller	Mary Davies	42	Female	Teacher	Panasonic HC-V10	6:56
7	Pablo.mpg	24.4.2012	Matt Miller	Pablo Neftali	76	Male	Poet	Canon XF305	4:32

Periodicals, magazines, journal articles

Among materials you use for qualitative data analysis, there may be online periodicals, magazines or journal articles. The information about all such resources must be kept in separate files:

- » Material collected from online periodicals: save references to web resources, like URLs, and do not forget they may change over time. To be sure information is not lost, articles should be copied into a word processing program;
- » Materials from periodicals: When articles, photographs and other material are collected from periodicals for research purposes, bibliographic information should be carefully detailed (author(s), title, date of publication etc.);
- » When you analyse articles, make a list of them, sort them alphabetically or chronologically in the order they were analysed in the course of research.

Storing documentation

- » Write the documentation into a separate, well-structured file, and associate that with the data file. You may use the same filename stem in order to strengthen the file-metadata association. For example: 20130311_interviews_audio, 20130311_interviews_trans, 20130311_interviews_image, 20130311_interviews_metadata. The latter part of the name can be used to convey the specifics of the file. In this case “audio” means audio tape and “trans” a transcription of the audio tape;
- » Data-level documentation can be embedded within a data file. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file;
- » If you have a large amount of metadata or large amounts of data that will need metadata you can use a standard specific database for this purpose (such as the DDI Codebook (DDI Alliance, 2017a)).

Metadata: machine readable data documentation

Metadata or “data about data” are descriptors that facilitate cataloguing data and data discovery. Metadata are intended for machine-reading. When data is submitted to a trusted data repository, the archive generates machine-readable metadata. Machine-readable metadata help to explain the purpose, origin, time, location, creator(s), terms of use, and access conditions of research data.



Create machine-readable metadata

Check out [The Dublin Core Metadata Generator](#) (dublincoregenerator, n.d.) and see how metadata elements are converted into a machine-readable file in *.xml.

Also, if you enjoy working with *.xml schemas, get started in creating a codebook to accompany your dataset with the [DDI codebook](#) (DDI Alliance, 2017a).

Deposit data in a data repository

When you submit your dataset in a (trusted) data repository, machine-readable metadata will be added. See the chapter on ‘[Archiving and Publishing data](#)’ for a description of such data repositories.

In the boxes below we provide you with examples of:

» **Metadata templates** (for easy starting)

If you do not quite know yet what metadata you should generate (what fields are needed) have a look at the metadata templates provided. Some of them are very simple and can, therefore, help to create basic documentation.

» **Metadata standards** (for when you need your metadata to be very structured).

Metadata standards may at first look seem quite scary. They are used by data archives for enhancing discoverability, interoperability and reusability. When you submit your dataset to a trusted data repository, these standards are automatically applied.

Metadata templates

Metadata can, at its simplest, be stored in a single text file. However, you can also use a metadata template to help you structure your metadata or to see how your metadata appears in *.html.

Below we provide examples of metadata templates that you can use when compiling documentation. Or just for inspiration to take a look at typical fields which are often required. It is always possible to include additional documentation beyond what is suggested.

- » [Create a codebook](#) about your research to accompany the dataset (DDI Alliance, 2017a).
- » Download the York University (n.d.) [Library Metadata Template](#), Dublin Core;
- » Have a look at the Georgia Tech Library (n.d.) [Metadata Template](#);
- » Use the [Dublin Core Metadata Generator](#) (dublincoregenerator, n.d.);
- » Have a look at the Cornell University (n.d.) [guide to writing "readme" style metadata](#) (with downloadable template);
- » Use the [ISO 19115-2 Metadata Editor](#) (GRIIDC (2015)) web application.

Metadata standards

You may want your metadata to be very structured. For that purpose, you can choose a metadata standard or a tool (software that has been developed to capture or store metadata) to help you add and organise your documentation. Many standards are discipline-specific. These will help you to add metadata to the workflow as they have been created to suit the needs of research data.

Remember that you do not generally need to generate machine-readable metadata by yourself. The repository where you may want to deposit your data will do that for you. When you are depositing your data the repository will require a data documentation document from you and will convert the documentation into machine-readable metadata.

The recommended standard for research in the social sciences is the DDI metadata standard.

DDI for social sciences

[DDI](#) (Data Documentation Initiative) (DDI Alliance, 2017b) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences. Expressed in XML, the DDI metadata specification supports the entire research data lifecycle.

Common fields in the DDI include:

- » Title
- » Alternate Title
- » Principal Investigator
- » Funding
- » Bibliographic Citation
- » Series Information
- » Summary
- » Subject Terms
- » Geographic Coverage
- » Time Period
- » Date of Collection
- » Unit of Observation
- » Universe
- » Data Type
- » Sampling
- » Weights
- » Mode of Collection
- » Response Rates
- » Extent of Processing
- » Restrictions
- » Version History

MIDAS Heritage for historical sites

[MIDAS Heritage](#) (Historic England, 2012) is a British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

VRA Core for images and works of art and culture

[VRA Core](#) (2015) is a standard for the description of images and works of art and culture.

ISO 19115 for geospatial data

[ISO 19115](#) (DCC, 2017) is a schema for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, the spatial reference, and the distribution of digital geographic data.

Metadata standards for general research data

- » [Dublin Core](#) (DCMI, 2017);
- » [DataCite Metadata Schema](#) (Datacite, n.d.);
- » [PREMIS](#) (2017).

In its simplest form, the Dublin Core consists of 15 fields that basically describe all online resources:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Example of DDI

For an example of how to apply the metadata standard DDI, have a look at [a dataset in the Finnish Social Science Data Archive](#) (Galanakis, Michail (University of Helsinki): Intercultural Urban Public Space in Toronto 2011-2013 [dataset]. Version 1.0 (2014-02-13). Finnish Social Science Data Archive [distributor]. <http://urn.fi/urn:nbn:fi:fsd:T-FSD2926>).

The machine-readable XML file [looks like this](#).

For a visually formatted example of a DDI record, see the online version of this chapter: <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata>

Metadata for new data types – new standards still under development

To provide metadata for social media data and transaction data with metadata, the metadata standards by the [Data Documentation Initiative](#) (DDI) should serve as the guiding framework. Importantly, however the DDI standard is “insufficient to document all the details required for reproducibility of a social media dataset” ([Kinder-Kurlanda et al 2017: 3](#)). For example, the DDI format does not allow describing biases caused by data mining interfaces of social media platforms, changes in data availability and formats, explanations about code and scripts used in collection, cleaning and analysis etc. Such information can be described only in an unstructured manner as an additional comment in the standard’s form.

Together with other CESSDA partners, GESIS is currently developing recommendations for the provision of metadata for new data types (esp. social media data).

2.4 Adapt your DMP: part 2



This is the second of six 'Adapt your DMP' sections in this tour guide. After working on this chapter, you should be able to plan for organising and documenting your data.

To adapt your DMP, consider the following elements and corresponding questions:

Document data type and size

- » What type(s) of data will be collected?
- » What is the scope, quantity, and format of the material?
- » What is the total amount of data collected (in MB/GB)?

Data organisation

- » How will you organise your data?
- » Will the data be organised in simple files or more complex databases?
- » What is your process for quality assurance? What are your quality measures?
- » Are there specific quality standards or quality management models you plan to apply?

File naming and folder structure

Are there any specific requirements for compatibility and comparability of your data?

Are there specific standards that you want to implement, e.g. naming conventions or standardised coding structures?

Data documentation and metadata

- » Will you be creating separate files accompanying the data?
- » Will you be using a database?
- » Are the data produced and/or used in the project discoverable with metadata?
- » What metadata will you use? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- » If you already know in which data repository you will publish your data, what metadata standard do they use?

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 3

Process

Contents

Main take-aways	60
3.1 Data entry and integrity.....	61
3.2 Quantitative coding.....	66
3.3 Qualitative coding.....	70
3.4 Weights of survey data	72
3.5 File formats and data conversion.....	76
3.6 Data authenticity	79
3.7 Wrap up: Data quality	82
3.8 Adapt your DMP: part 3	84
Sources and further reading	85

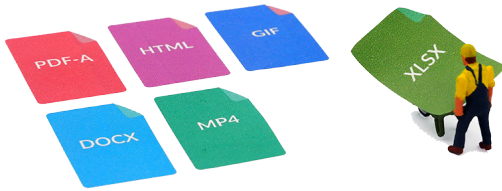
[View the online version of this chapter](#)

Main authors of this chapter

Jindrich Krejčí, Czech Social Science Data Archive (CSDA)

Johana Chylikova, Czech Social Science Data Archive (CSDA)

Introduction



In this chapter¹, we focus on the data operations needed to prepare your data files for analysis and data sharing. Throughout the different phases of your project, your data files will be edited numerous times. During this process, it is crucial to maintain the authenticity of research information contained in the data and prevent it from loss or deterioration.

However, we will start with the topics of data entry and coding as the first steps of your work with your data files. Finally, you will learn about the importance of a comprehensive approach to data quality.

Main take-aways

After completing your journey through this chapter on organising and documenting your data you should:

- » Be familiar with strategies to minimise errors during the processes of data entry and data coding;
- » Understand why the choice of file format should be planned carefully;
- » Be able to manage the integrity and authenticity of your data during the research process;
- » Understand the importance of a systematic approach to data quality;
- » Be able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt them to your own DMP

¹ The content of this chapter was inspired by research data management manuals, guidelines, online courses and methodological texts published by several data organisations and experts, in particular the [information provided by the UK Data Service](#) (2017a), the “[Guide to Social Science Data Preparation and Archiving](#)” by the US-based data organisation ICPSR (2012), the online course [Research Data MANTRA](#) (EDINA and Data Library, University of Edinburgh, 2017), A guide into research data management by Corti, Van den Eynden, Bishop and Woollard (2014), Krejčí’s “Introduction to the Management of Social Survey Data” (Krejčí, 2014), Gibbs (2007) and [Data Management Guidelines](#) produced and published by the Finnish Social Science Data Archive (Finnish Social Science data Archive, 2017).

3.1 Data entry and integrity

Data integrity means assurance of the accuracy, consistency, and completeness of original information contained in the data. At the same time, the authenticity of the original research information has to be preserved (see '[Data authenticity](#)').

The integrity of a data file is based on its structure and on links between data and integrated elements of documentation. From the moment that data is being entered, data integrity is at stake.

Data entry procedures have changed over recent years. Operators entering data into a computer manually are being replaced by automated computer technologies, while the universal distinction between the three phases of data collection, data entry, and data editing/checking is often becoming obsolete. In general, greater automation of processes generally prevents some types of errors, but at the same time, it produces other types of errors. For example, errors in scripts during computer-assisted interviewing may cause systematic shifts in data and to be able to detect such deviations in automated forms of data entry requires different kinds of checks in comparison to manually entered data.

Minimising errors in survey data entry

In the boxes below, a summary of recommendations on minimising errors in survey data entry is given (UK Data Service, 2017a; ICPSR, 2012; Groves et al., 2004).

Check the completeness of records

Check if your data files contain the correct number of records, number of variables or length of the records, etc.

Reduce burden of manual data entry

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.

Minimise the number of steps

The data entry process should include a smaller rather than a larger number of steps. This reduces the likelihood of errors.

Conduct data entry twice

When you have paper questionnaires, the data entry can be processed electronically by scanning questionnaires or manually entering data by a person responsible for data entry. If data are entered by scanning, execute the process of data entry twice and compare values. If data are entered manually, a portion of questionnaires should be entered twice by two different persons. For example, the Czech Association of Public Opinion and Market Research Agencies ([SIMAR](#)) recommends 20 percent of questionnaires be re-entered.

Perform in-depth checks for selected records

At least some randomly selected records, e.g. 5–10% of all records, should be subjected to a more detailed, in-depth check to verify the procedures and identify possible systematic errors. The cases should be selected by chance. Be sure to document the changes you make and keep the original data so you can restore them at all times.

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 the available time);
- » Compare your data with historical data (e.g. check the number of household members with the previous wave of a panel survey).

Automate checks whenever possible

Specialised software for computer-assisted interviewing (CAPI, CATI, etc.) or data entry software allows to set the range of valid values for each category and to apply filters to manage the data entry or the entire data collection process. These automatic checks:

- » Prevent meaningless values from being entered;
- » Help to discover inconsistencies that arise when some values are skipped or omitted;
- » Make the interviewer's work substantially clearer and easier;
- » Reduce the number of errors that interviewers make.

The software can distinguish between permanent rules that cannot be bent and warnings that only notify the operator when entering an unlikely value.

CAPI software is used by the data collectors and it is usually expensive and therefore individual researchers cannot afford to buy it. In case you collected your survey data by yourself, you must write your own program/syntax to check your data for discrepancies.

An example of an SPSS syntax to check your data

Logical check of income - the household income cannot be SMALLER than individual income

The syntax search for respondents who indicated the household income as well as their individual income, while the household income was smaller than individual income.

Variable names:

ide.10 - household income

interval variable, income in Euros, with special values 8 - refused to answer; 9 - don't know

ide.10a - individual income

interval variable, income in Euros, with special values 8 - refused to answer; 7 - doesn't have income

Syntax (SPSS):

USE ALL.

```
COMPUTE filter_$=(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and (ide.10 ne 8) and (ide.10 ne 9) and (ide.10 < ide.10a).
```

```
VARIABLE LABELS filter_$ '(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and (ide.10 ne 8) and (ide.10 ne 9) and (ide.10 < ide.10a) (FILTER)'.  
/FILTER.
```

```
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
```

```
FORMATS filter_$ (f1.0).
```

```
FILTER BY filter_$.
```

```
EXECUTE.
```

```
FREQUENCIES VARIABLES=CD
```

```
/ORDER=ANALYSIS.
```

```
FILTER OFF.
```

USE ALL.

In cases of errors ...

What to do with error values?

You 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, you should bear in mind that inconsistencies can also be generated by the respondents themselves, and a correction should make a minimum or no changes/reductions to their original answers. Any replacement of originally measured values must be planned for and done in conformity with your research concepts.

Entering data directly into the MS Excel sheet or data list sheets of statistical software packages is a source of frequent errors. It is easy to skip the column or row and then it is difficult to identify all the errors and correct them. However, even in MS Excel, it is easy to set up a form for purposes of entering the records one by one ([see the video by United computers, 2013](#)) video and set up some simple checks if you have at least basic programming skills. Using MS Access for this purpose would be easier. It is also possible to use suitable data entry freeware, which is widely available from the web.

Considerations in making high-quality transcriptions of qualitative data

The most common formats of qualitative data are written texts, interview data and focus group discussion data. In most cases, interview and discussion data are firstly digitally recorded and then transcribed. Transcription is a translation between forms of qualitative data, most commonly a conversion of audio or video recordings into text. If you intend to share your data with other researchers, you should prepare a full transcription of your recordings (Bucholtz, 2000).

There are several basic rules and steps in the process of making and checking a high-quality transcript from audio/video (Kuckartz, 2014):

Prevent mistranscription by recording high-quality data

The quality of interview data gathered by means of recorded interviews depends on both the skills of the interviewer and the quality of the audio-visual equipment. Taking steps to create audio recordings of good quality increases their usefulness. Good quality sound recordings should prevent mis-transcription and reduce the chance of sections of an interview remaining untranscribed due to poor sound quality. When recording an interview, consider the following (Bucholtz, 2000):

- » The level of sound or picture quality needed;
- » The budget available for equipment and related consumables;
- » How quickly the technology being used will become redundant;
- » Whether consent is in place to allow the fullest use of recordings;
- » How the data created will be used;
- » Whether data or information not allowed by consent can be excluded from recording;
- » Whether the equipment will be simple to operate in the field.

Determine the transcription method

Transcription methods depend upon your theoretical and methodological approach and can vary between disciplines. Three basic approaches to transcription are (Bucholtz, 2000):

- » **Focus on the content**
This is also called the denaturalised approach, most like written language. The focus is on the content of what was said and the themes that emerge from that. This approach is used in sociological research projects.
- » **Focus on what is said and how it is said**
This approach is called the naturalised approach, which is most closely to speech. A transcriber seeks to capture all the sounds they hear and use a range of symbols to represent particular features of speech like the length of pauses, laughter, overlapping speech, turn-taking or intonation. This approach is usually employed in projects using conversation analysis.
- » **Focus on emotional and physical language**
In this approach detailed notes on emotional reactions, physical orientation, body language, use of space, as well as the psycho-dynamics in the relationship between the interviewer and interviewee are detailed. This approach is usually used in psycho-social research.

Choose between manually transcribing or with the help of speech recognition software (SRS)

SRS must “get used” to a speaker and can only be used if a high-quality recording is available. Gibbs (2007) recommends checking the utility and functionality of SRS software before using it.

When transcribing manually, you may sometimes hear something other than what an interviewee actually said. Listen carefully.

Determine the rules

- » Determine a set of transcription rules or choose an established transcription system that is suited for the planned analysis;
- » In setting up the rules, consider compatibility with the import features of QDA (Quality Data Analysis) software. For example, document headers and textual formatting, such as italics or bold, may be lost when transcripts are imported into software packages, and text formatted in two columns indicating speakers and utterances may also be problematic;
- » All members who are doing the transcription should first agree on these rules;
- » Write transcriber instructions or guidelines with required transcription style, layout and editing.

Transcribe

Transcribe the texts (or part of the texts) on the computer.

Check the transcription

Proofread, edit and modify the transcription, if necessary.

Protect your participants

- » Anonymise data during transcription, or mark sensitive information for later anonymisation (see ['Anonymisation'](#));
- » When you assign the task of transcription to somebody else, make sure to take care of personal data protection before sending audio recordings and transcripts that contain personal or sensitive information. Draw up a non-disclosure agreement with the transcriber and encrypt files before transfer.

Choose a QDA-compatible file format

Format the transcription in such a way that your QDA (Qualitative Data Analysis) can be used optimally and files can be imported into the QDA software.

Choose a file format for long-term preservation

Save and archive the transcription in long-term preservation ready files such as *.rtf or *.pdf files (see ['File Formats'](#)).

3.2 Quantitative coding

Quantitative coding is the process of categorising the collected non-numerical information into groups and assigning the numerical codes to these groups. Numeric coding is shared by all statistical software and among others, it facilitates data conversion and measurement comparisons.

Closed-ended questions

For closed-ended questions in survey questionnaires, the coding scheme is often incorporated directly into the questionnaire and data is entered numerically. This process is automated in computer-assisted interviewing (CAPI, CATI, etc.), where an answer and its code are saved immediately into a computer in the course of data collection. Answers can also be coded on paper questionnaires when coders record codes in a designed spot of the questionnaire before they are digitalised. If the numerical codes are not incorporated in your questionnaire, set up a detailed procedure of how to code the different alternatives.

Open-ended questions and other textual information

More complex coding exercises, e.g. for textual answers in survey questionnaires, 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.

Documentation

The meaning of codes must be documented. Specialized analytic software (SPSS, SAS, STATA, etc.) lets the user assign labels directly to the codes. For the principles of the construction of labels, please, see the sub-section '[Organising variables](#)'. If the software does not allow you to assign code labels directly to data, you have to document the codes in a separate document as part of the metadata.

Coding recommendations

In the boxes below you find coding recommendations which are inspired by ICPSR (2012).

Include identification variables

All identification variables should be included at the beginning of your data file. Identification variables usually include a unique identification of your study/data file, unique ID numbers of cases in your data file (e.g. ID of the respondent, ID of his/her household, etc.) as well as the identification of other characteristics essential for analysis (e.g. identification of different methods of data collection or sources, identification of the over-sample, etc.).

Make code categories exclusive and coherent throughout the database

Code categories should be mutually exclusive, exhaustive, and precisely defined. Ambiguity will cause coding difficulties and problems with the interpretation of the data. You should be able to assign each response of the respondent into one and only one category.

Preserve original information

Recording original data, such as age and income, is more useful than collapsing or bracketing the information. With original or detailed data, secondary analysts can determine other meaningful brackets on their own rather than being restricted to those chosen by others.

Document the coding schemes

Responses to closed-ended questions should retain the original coding scheme to avoid errors and confusion. For open-ended questions, investigators can either use a predetermined coding scheme or construct a coding scheme based on major categories that emerge in survey responses. Any coding scheme and its derivation should be reported in study documentation.

Check verbatim text data for data disclosure risk

Responses recorded as full verbatim (word for word) must be reviewed for disclosure risk and if necessary treated in accordance with applicable personal data protection

Check coding

It is advisable to verify the coding of selected cases by repeating the process with an independent coder. This provides means for verification of both the coder's work and the functionality of your coding scheme.

Distinguishing between major and lower level categories

If a series of responses require more than one field or if the response is very complex (for example a detailed description of one's occupation), it is advisable to apply a coding scheme distinguishing between major, secondary and any possible lower level categories. The first digit of the code identifies a major category, the second digit can distinguish specific responses within the major categories, etc.

[The International Standard Classification of Occupations](#) (ISCO) (International Labour Organisation, 2016) is an example of such a hierarchical category scheme. An example of its use is given below.

Standardised coding schemes

The use of standardised classifications and coding schemes brings many advantages, e.g.:

- » Economic and quality benefits as a result of adopting an existing structure which has a solid basis and has been verified in many studies;
- » Comparability with data from other studies using the same concept;
- » Comprehensibility for researchers who work with these concepts.

A disadvantage lies in the necessity to adapt your research intentions in line with the concept of the coding scheme.

Several standardised classification and coding schemes exist that you can use. For coding occupations it is the [International Standard Classification of Occupations](#) (ISCO) (International Labour Organisation, 2016), for coding education it is the [International Standard Classification of Education](#) (ISCED) (Unesco, 2011), for geographic territories it is the [Nomenclature of territorial units for statistics](#) (NUTS) (Eurostat, 2013), for economic activities it is the [Statistical classification of economic activities](#) (NACE) (Eurostat, 2008), for languages it is [ISO 639.2](#) (Library of Congress, n.d.), for disease it is the [International Classification of Diseases](#) (ICD) (World Health Organisation, 2016), etc.

Example

Occupational classifications such as such as the [International Standard Classification of Occupations](#) (ISCO) (International Labour Organization, 2010) are examples of widespread standard coding schemes. ISCO is an example of a hierarchical category scheme.

Occupational information has several dimensions and in questionnaire surveys, these need to be collected in detail. This is, as a rule, done by means of one or more open-ended questions.

The current ISCO-2008 uses four-digit codes. In the table below you see some examples.

- » 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](#) (2016).

For an example of a recommended methodology of collection of information on occupations see [Ganzeboom](#) (2010).

Coding missing values

Not all the questions in a questionnaire are answered by all respondents, which results in missing values on a variable level in the data file (so-called item non-response). It is crucial for data integrity to distinguish at least the situations when values are missing, because the variable is not applicable to the particular respondents.

Furthermore, it is often useful for analyses to identify whether the value is missing because the respondent did not know the answer, refused to answer or simply did not answer or consider other reasons for missing values (see the example below). The information on missing values is always an important part of your documentation and promotes transparency of your research work. However, bear in mind that possibilities to differentiate between many different types of the missing values in analysis can be limited by the abilities of your software.

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 or 97, 98, 99 or 997, 998, 999, etc. (where the number of digits corresponds to the variable's format and the number of valid values) are used for numeric coding of missing values. The coding scheme for missing values 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.

Example

Respondents in surveys sometimes do not answer all questions in a questionnaire. It is advisable to distinguish between various reasons that data went missing (ICPSR, 2012). The following situations are distinguished in survey research (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.

Training coders to prevent coder variance

Coders may vary in the way they assign codes to variable values, i.e. each of them uses the same coding scheme in a slightly different way. This results in so-called "coder variance". Coder variance is a specific source of non-sampling error (i.e., error additional to the statistical "sampling" error) and may cause systematic deviations of the sample.

Coding of textual information is a complicated cognitive process and the coder may pose a significant influence on the information that appears in the database, as well as become a source of systematic error. That is why the implementation of complicated coding schemes often requires the construction of a theoretically and technically well-founded design and requires specific coder's competencies and training.

3.3 Qualitative coding

Coding is a way of indexing or categorizing the text in order to establish a framework of thematic ideas about it | Gibbs (2007).

In qualitative research, coding is “how you define what the data you are analysing are about” (Gibbs, 2007). Coding is a process of identifying a passage in the text or other data items (photograph, image), searching and identifying concepts and finding relations between them. Therefore, coding is not just labeling; it is linking of data to the research idea and back to other data...

The codes which are applied enable you to organise data so you can examine and analyse them in a structured way, e.g. by examining relationships between codes.

Approaches to coding qualitative data

A basic division between coding approaches is concept-driven coding versus data-driven coding (or open coding). You may approach the data with a developed system of codes and look for concepts/ideas in the text (concept-driven approach) or you can look for ideas/concepts in the text without a preceding conceptualisation and let the text speak for itself (data-driven coding). Investigators can either use a predetermined coding scheme or review the initial responses or observations to construct a coding scheme based on major categories that emerge.

Both methods require initial and thorough readings of your data and writing down which patterns or themes you notice. A researcher usually identifies several passages of the text that share the same code, i.e. an expression for a shared concept.

An example

A code in a qualitative inquiry is most often a word or short phrase. In the table below an example (Saldaña, 2013) is given.

Raw data	Preliminary codes	Final code
The closer I get to “retirement age” the faster I want it to happen. I’m not even 55 yet and I would give anything to retire now. But there’s a mortgage to pay off and still a lot more to sock away in savings before I can even think of it. I keep playing the lottery, though, in hopes of dreams of early winning those millions. No retirement luck yet.	* retirement age* financial obligations dreams of early retirement	RETIREMENT ANXIETY

Expert tips



Any researcher who wishes to become proficient at doing qualitative analysis must learn to code well and easily. The excellence of the research rests in large part on the excellence of the coding | Strauss (1987).

Tip 1: Document the meaning of codes

The meaning of codes must be documented in a separate file. Make short descriptions of the meaning of each code. It is helpful to you and also to other researchers who will have access to your data/analysis. What you need to know about your codes (Gibbs 2007):

- » the label or name of the code
- » who coded it (name of the researcher/coder)
- » the date when the coding was done/changed
- » definition of the code; a description of the concept it refers to
- » information about the relationship of the code to other codes you are working with during the analysis.

Tip 2: Prevent coder variance

Coding textual information is a complicated cognitive process and the coder is necessarily a significant influence on the coding process. For each study coding procedures must be carefully planned and a specific coding design and guidelines must be established. Coders must undertake a training, where they are instructed about the specific coding design and coding rules. A part of coding procedures is concerned with reviewing the quality of the coding process. According to Gibbs (2007) several techniques to control coder reliability exist:

- » **Checking the transcription**
An independent researcher goes through coded texts and considers the degree to which coders differed from each other.
- » **Checking for definitional drift in coding**
If you code a large dataset the data at the beginning may be coded slightly different than material coded later. Check the the whole dataset for the definitional drift. Have good notes with descriptions of individual codes.
- » **Working in a team**
If there are multiple people working in a team, individual members can check each other´s coding.

3.4 Weights of survey data

When conducting a survey, having a representative sample of the population is of paramount importance. But in practice, you are prone to over-sample some kinds of people and under-sample others. Weighting is a statistical technique to compensate for this type of 'sampling bias'. A weight is assigned to:

- » Reflect the data item's relative importance based on the objective of the data collection;
- » Take into account the characteristics of sampling design;
- » Reduce bias arising from nonresponse when the characteristics of the respondents differ from those not responding;
- » 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.

Different types of weights and their different purposes

Several types of weights have different purposes and a different impact on data analysis.

An answer to the question whether or not to use weights is not straightforward. For particular methods of analysis (e.g., estimating associations, regressions, etc.) using weights may be dysfunctional. There are also general theoretical and methodological issues which discourage some researchers from using weights. However, different types of weights are useful for different purposes. In some situations, it is necessary to take an appropriate weight into account in your analysis (see several types of weighting below).

In all cases, if there are any weights in your data file, the rationale and calculation of the weights must be detailed in the data documentation.

Design weights

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.

To solve these differences in sampling probabilities we have to compute design weights. The design weights are equal to the inverse of the probability of inclusion in the sample. The sum of all design weights should be equal to the total number of units in our population.

Non-response weighting

During the implementation of a survey, we are normally not able to get a response from some of the targeted respondents we sampled due to:

- » Their refusal;
- » Our failure to contact them;
- » Other administrative reasons.

Response rates differ between various population groups and those inequalities can be compensated for by weighting.

Post-stratification weighting

The way certain characteristics such as sex, age, and education of your sample population are distributed may differ from the way it is distributed in the actual population. For example, your sample may consist of 66 percent men when they make up only 48 percent of the population. Post-stratification weighting is done in order to achieve a distribution equal with that of such known characteristics of the population. It is called a post-stratification weight because it can only be computed after you have collected all of your data. Stratification comes from the various known strata (such as age group or sex distribution) of the population.

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

The data file may include several different types of weights for different purposes. Subsequently, they are combined into a final, combined weight.

An example: Comparison of weighted and non-weighted data

Source: Data files from the [ESS, round 8, Czech Republic](#) (European Social Survey, 2016).

Variable name: netusoft

Question: How often a respondent uses internet

In the first column, no weight was applied.

In the second column, the Design Weights (DWEIGHT) are adjusted for different selection probabilities.

	No weight		Design weight	
	Frequency	Valid Percent	Frequency	Valid Percent
1 Never	244	10,8	187	8,2
2 Only occasionally	162	7,1	155	6,8
3 A few times a week	302	13,3	284	12,5
4 Most days	384	16,9	379	16,6
5 Every day	1177	51,9	1271	55,8
Total	2269	100	2277	100
System missing	31		23	
Total	2300		2300	

Distribution of weights

If the weight of a case equals 1 then the values measured are not adjusted. In the case of post-stratification weights both high or low numbers indicate either large deviations of the sample from the target population, poor quality of the weight or both. It is desirable that the large part of values of the weighting variable is close to 1.

Weights constructed by others

Is there any weighting variable in your working data file? If yes and you are not the author of the weight, never use it without knowledge of its origin and purpose. You should always thoroughly explore the distribution of the weighting variable and its impact on distributions of other selected variables from the data file.

An example: Using weights in European Social Survey data

The following table provides an illustration of using weights in the data from the [European Social Survey](#) (n.d.) (ESS). There are three different weights available in the ESS Source Main Questionnaire data file (see [European Social Survey, 2014](#)):

1. The design weight takes into consideration the different probabilities of being sampled given the sampling methods implemented in individual countries;
2. The post-stratification weight corrects for the differences of the sample from selected population characteristics caused by other sampling and non-sampling errors;
3. The population size weight corrects the fact that the individual countries' sample sizes are very similar while there are large variations in the size of their actual populations.

Different types of data analysis then require the use of different weights or their combinations. When analysing data from one country alone or comparing data of two or more countries, only the design weight or the post-stratification weight needs to be applied. When combining different countries, design or post-stratification weights in combination with population size weights should be applied.

	Example – voter turnout (% of respondents voting in the last election)	Weights to be used	
		Design weight / Post-stratification weight	Population weight
To examine data from a single country – whether a single variable or a cross-tabulation	Voter turnout in Germany	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 cross-tabulation	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, 2014.*

3.5 File formats and data conversion

We use software for creating text documents, websites, databases, photos, 3D models, and movies. Software developers regularly release new versions of their products. It is not self-evident that the new software supports the use of files created with earlier software versions (compatibility). And some software packages even disappear completely from the scene. Conversions of file formats may be costly or result in loss of information or a reduction of data quality. This is exactly why the choice of file formats should be planned carefully.

Short-term data processing: file formats for operability

File format choice depends on your research phase. Choices for short-term data processing may differ from the choices you make for long-term data preservation.

For the reasons of short-term operability, it is advisable to choose a file format that is associated with the specific software that you intend to use for data analysis. Following discipline-specific standards and customs is generally the way to go. However, you should take into consideration how widespread these standards are and to what extent they will allow data processing by others than peers in your own discipline.

Proprietary file formats are owned and copyrighted by a specific company. Their specifications are usually not publicly available and their future development results from decisions and situation of their owner. Thus, the risk of obsolescence is high. However, some proprietary formats, such as Rich Text Format (*.rtf), MP3, MPEG, JPG, MS Excel (*.xls), SPSS (*.sav, *.por), STATA (*.dta) are widely used and you may assume that they will be useful for a reasonable time.

Learn more about suitable file formats for short-term data processing

Below we give an overview of the data analysis packages/file formats which are used most and which are suitable for short-term data processing.

Quantitative (statistical) data analysis packages

MS Excel (*.xls), SPSS (*.sav, *.por), R and STATA (*.dta) are widely used and you may assume that they will be useful for a reasonable time.

Some software also provides so-called portable formats which allow easy transfer of data between different versions of the software of the same brand, often including versions for different platforms (MS Windows, Mac, Linux...). For example, SPSS system files with the *.sav extension and SAS files with the *.sd7 extension (SAS Version 7 or 8 data file) are associated with the concrete version of the SPSS or SAS software. Instead of them, you may use "portable" SPSS files with the *.por extension or "transport" SAS files, which are compatible with different versions of this software running on different platforms.

Qualitative data analysis packages

Qualitative research data like transcribed interviews of focus group sessions, audio recordings, still images, photographs, ethnographic diaries and various types of written texts are usually transcribed into one of the following types of formats: *.docx, *.rtf, *.pdf, *.mp3, *.wav, *.jpeg and many others.

For the purposes of qualitative data analysis (QDA), textual data may be analyzed in special QDA software packages such as NVivo, ATLAS-ti, and MAXQDA. In such packages researchers are allowed to code their textual data, i.e. indicate parts of text related to same concepts, create a structure of concepts etc. In the process of coding, a "coding tree" emerges along other pieces of information, for example, notes and memos. Common QDA packages have export facilities that enable a whole 'project' consisting of the raw data, coding tree, coded data (Also see '[Coding qualitative data](#)'), and associated memos and notes to be saved.

Long-term data preservation: file formats for the future

Standard, open and widespread formats are advisable for long-term storage as they typically undergo fewer changes. Contrary to proprietary formats (see above) specification of open formats is publicly available. Some of them are standardised and maintained by a standards organisation and we may assume that their readability in the future is ensured. Examples of open formats are PDF/A, CSV, TIFF, ASCII, Open Document Format (ODF), XML, Office Open XML, JPEG 2000, PNG, SVG, HTML, XHTML, RSS, CSS, etc.

Learn more about file formats for long-term preservation

Quantitative data preservation

Long-term preservation of quantitative data is typically best off with simple text (ASCII) formats accompanied by a structured documentation file with information about the variables included, their position in the file, formats, variable labels, value labels etc.

In terms of location of variables in the file, we distinguish between fixed and free formats.

Fixed format In a fixed format, variables are arranged in columns and their exact positions, i.e. the start and end of each variable, are known.

Free format In a free format data for each variable is separated by blanks or specific characters, e.g. by tab space or a dash. If the character separating variables is used within an item, then it needs to be formatted specifically and separated from the surrounding text (as a rule, by quotation marks).

There exist several extensions for simple text formats, e.g. *.txt., *.dat and *.asc are used for both fixed and free formats, *.csv. is used for fixed format.

Qualitative data preservation

Qualitative data analysis software packages such as NVivo, ATLAS-ti, and MAXQDA have export facilities that enable a whole 'project' consisting of the raw data, coding tree, coded data, and associated memos and notes to be saved. For archiving such data, the raw data, the final coding tree, and any useful memos should be exported (UK Data Service, 2017)

Digital versions of documents are usually kept in the PDF/A format. This is an official archiving version of the PDF format as defined by the ISO 19005-1:2005 standard. It guarantees independence from the platform and includes all display information (including fonts, colours, etc.). XMLP format is a widespread standard for metadata. Structured textual documentation should, again, be saved in a simple text format, with tags and in line with a standard structure (e.g., DDI).

For audio files the recommended longterm format is WAV, video files are advised to be stored in MXF (Material eXchange Format) and JPEG2000 ([Fleischhauer, 2010](#)).

A very useful tool for searching an appropriate format for different types of data is provided by the UK Data Service (2017b) in the [table of Recommended file formats](#).

Data conversion and possible data loss

Data files, depending on the nature of the data, are based on either text or binary encoding or both. Binary encoded information can be read only by specialised software, text information is universal and can be read by a wide range of different software including text editors.

It is advisable to store your data for use in the future, which means converting them from a current data format to a long-term preservation format. Most software applications offer export or exchange formats that allow a text-formatted file to be created for importing into another program. A typical example is Microsoft Excel, which through the 'Save As' command, can save spreadsheet data in comma delimited format (*.csv or comma separated values). The structure of the rows and columns is preserved through commas and line returns. However, multiple worksheets must be saved as separate *.csv files and any text formatting or macros in the native format will be lost on conversion.

During the process of data conversion, important pieces of information may be lost:

- » In the conversion of a statistical dataset (i.e. survey data), parts of the dataset may be lost, same as missing data definitions, decimal numbers, changes in data formats (e.g., numerical into string data type), data also may be truncated;
- » In case of texts, i.e. transcriptions of speech, editing such as highlighting, bold texts, headers, footers may be lost;
- » In case of images a reduction of resolution, loss of layer, colours may be lost;
- » In converting audiovisual data file conversion may reduce sound quality;
- » Some file formats are constructed specifically to save space. However, this is done by a reduction of information and data quality. For example, .jpg removes details from images, while .tiff bears full information. Similarly, .mp3 is a lossy format for audio data, while .wav keeps detailed information.

For this reason, the conversion itself should be done by a researcher familiar with the data, so he or she can check for potential undesirable changes in the data that occurred as a result of the conversion.

Due to differences in national character sets you should pay attention also to character coding. Some coding systems (e.g., Windows 1250) do not cover all character sets at the same time. As a result, an adequate language environment (Central European languages) has to be set to ensure correct display, which cannot be done at all times. Other coding systems (e.g., UTF 8) allow correct display of symbols of several character sets simultaneously.

TIP: Plan ahead to simplify data publication



Different data archives have different preferred formats. Knowing about these preferred formats in advance can save you time later when you want to archive and publish your data. Usually preferred formats are frequently used, independent of specific software, and have open specifications (see for instance information by [DANS](#) (n.d.) on preferred formats).

3.6 Data authenticity

Processing and analysis of data inevitably result in a number of edits in the data file. However, it is necessary to preserve the authenticity of the original research information contained in the data throughout the whole data lifecycle.

There are many possible types of changes in the data:

- » Data cleaning procedures may be implemented;
- » Errors are often found and corrected;
- » New variables may be constructed;
- » New information may be added from external sources;
- » File formats may be changed;
- » New data may be included;
- » The data file structure may be changed for the purpose of increasing operability, etc.

As a result of above-mentioned data management processes, several different **versions** of the data file are usually created. They are important, as they allow you to step back to versions before particular changes were made. Versions may be used simultaneously for different purposes or replace one another. When data files are being published to make them widely available, the treatment of errors, inclusion of new data and/or changes in a data file structure may result also in the publication of new **editions** of the same data file which may substantially differ in their content (e.g. when new country data are included into an international data file).

Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

- » Clearly distinguish between individual versions and editions and keep track of their differences;
- » Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices

The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- » Establish the terms and conditions of data use and make them known to team members and other users;
- » 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;
- » Distinguish between versions shared by researchers and working versions of individuals;
- » Decide how many versions of a file to keep, which versions to keep (e.g. major versions rather than minor versions (keep version 02-00 but not 02-01)), for how long and how to organise versions;
- » Introduce clear and systematic naming of data file versions and editions;
- » Record relationships between items where needed, for example between code and the data file it is run against, between data file and related documentation or metadata or between multiple files;
- » Document which changes were made in any version;
- » Keep original versions of data files, or keep documentation that allows the reconstruction of original files;
- » Track the location of files if they are stored in a variety of locations;
- » Regularly synchronise files in different locations, such as using [MS SyncToy](#) (2016).

Version control

Version control can be done through:

- » Uniquely identifying different versions of files using a systematic naming convention, such as using version numbers or dates (date format should be YYYY-MM-DD, see ['File naming'](#));
- » Record the date within the file, for example, 20010911_Video_Twintowers;
- » Process the version numbering into the file name, for example, HealthTest-00-02 or HealthTest_v2;
- » **Do not** use ambiguous descriptions for the version you are working on. Who will know whether MyThesisFinal.doc, MyThesisLastOne.doc or another file is really the final version?
- » Using version control facilities within the software you use;
- » Using versioning software like [Subversion](#) (2017);
- » Using file-sharing services with incorporated version control (but remember that using commercial cloud services such as the Google cloud platform, Dropbox or iCloud comes with specific rules set by the provider of these services. Private companies have their own terms of use which applies for example to copyrights);
- » Designing and using a version control table. In all cases, a file history table should be included within a file. In this file, you can keep track of versions and details of the changes which were made. On the following page is [an example which was taken from the UK Data Service](#) (2017c).

Example of a version control table

Title:	Vision screening tests in Essex nurseries
File Name:	VisionScreenResults_00_05
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007
Created By:	Chris Wilkinson
Maintained By:	Sally Watsley
Created:	04/07/2007
Last Modified:	25/11/2007
Based on:	VisionScreenDatabaseDesign_02_00

Version	Responsible	Notes	Last amended
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Versioning new data types

Generally, the goal of version management is to enable reproducibility and support trustworthiness by allowing all transformations in the data to be traced. But difficulties emerge connected to versioning of “[new data](#)” as these data are (compared to “traditional data”) more frequently or even continuously updated. A good example are collections of Tweets (e.g., for a certain hashtag) as individual posts may be modified or deleted. As the contents of these data are continuously changing and if archived data are expected to reflect such changes (e. g. deleting posts from data set if they were deleted from platform) the result is an increasing number of versions. Consequently, it is necessary to develop a systematic plan to create and name new versions of constantly changing datasets, or find new solutions for streaming data.

Both researchers and repositories can learn from the fields where versioning of dynamic data is already established, such as the field of software development. The most common version control software in software development is Git. Some of the established repositories, such as [Zenodo](#) and [FigShare](#) or the [Open Science Framework](#), now offer integration with [GitHub](#), so that every version of data sets in those repositories can be recorded through it. A new project called [Dolt](#) is developing version control specifically for data which is particularly interesting for dynamic data sets, such as social media data.

To identify the exact version of a dataset as it was used in a specific project or publication, the [Research Data Alliance](#) (RDA) suggests that every dataset is versioned, timestamped, and assigned a persistent identifier (PID). In the case of Big Data, however, the [RDA warns](#) against excessive versioning: *“In large data scenarios, storing all revisions of each record might not be a valid approach. Therefore in our framework, we define a record to be relevant in terms of reproducibility, if and only if it has been accessed and used in a data set. Thus, high-frequency updates that were not ever read might go - from a data citation perspective - unversioned.”*

3.7 Wrap up: Data quality

The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise | American Association for Public Opinion Research (2015) (AAPOR).

How you organise, document and process data has a clear impact on data quality and thus also on reliability and adequacy of research findings. In scientific research even small things matter. Data organisation, documentation and processing procedures are not an exception and this is true for quantitative as well as qualitative research. A systematic approach and punctuality in data management (Krejčí, 2010) are awarded by the following:

- » Preventing errors and false findings;
- » Smooth course, time efficiency and transparency of your own research work;
- » Establishing assumptions for effective re-use of research data outside of the original research team.

While in quantitative research the quality is closely linked to standardization and control over the research situation the prevailing approach in qualitative research is different.

In qualitative research, discussions about quality in research are not so much based on the idea of standardization and control, as this seems incompatible with many qualitative methods. Quality is rather seen as an issue of how to manage it. Sometimes it is linked to rigour in applying a certain method, but more often to the soundness of the research as a whole | Flick (2007).

A complex approach to data quality

In previous chapters, you have become familiar with a number of procedures and rules for the development of an appropriate data file structure, development of rich metadata and ensuring the data integrity and authenticity. At the same time, however, we should bear in mind that the data management is always an integral part of much more complex research work.

The quality of the outcome is achieved through the quality of the production process (Krejčí, 2010). Scientific research is not an exception. Thus the quality stems from professionalism based on continuous improvement. Data management is one important part of such processes. As such it is interconnected and influenced by other processes within the system and should contribute to a common long-term objective of continuous improvement of a research work within the research organisation.

The mechanical quality control of survey operations such as coding and keying does not easily lend itself to continuous improvement. Rather, it must be complemented with feedback and learning where the survey workers themselves are part of an improvement process | Biemer & Lyberg (2003).

In addition, quality always involves a number of different dimensions. Quality is often defined as “fitness to use”. However simple this sounds, it provides a point of departure for a comprehensive approach to data quality. The results must not only be accurate but must be delivered in time, understandable and clear, and meet other potential users’ needs, e.g. comparability and coherence with other databases. Moreover, it must be also cost-efficient.

See the section on the next page for an example of how total quality management is handled by the [European Statistical System](#) (Eurostat, 2017).

Total Quality Management of the European Statistical System (ESS)

So-called models of Total Quality Management (TQM) recognise multiple dimensions of quality. They:

- » Set the required characteristics of a final product;
- » Define partial goals;
- » Elaborate the individual dependable processes to achieve them;
- » Identify and treat problematic points;
- » Specify control points;
- » Have procedures for quality monitoring, learning processes, and feedback-loops for ensuring continuous improvement.

Some of the guiding principles on which the [Quality Assurance Framework of the European Statistical System](#) (n.d.) is based are stated in the table below.

Data management procedures are an important part of such TQM models, building on similar principles and having the same goals.

Cost-effectiveness	Resources are used effectively.
Relevance	European Statistics meet the needs of users.
Timeliness and Punctuality	European Statistics are released in a timely and punctual manner.
Accuracy and Reliability	Source data, intermediate results, and statistical outputs are regularly assessed and validated.
Coherence and Comparability	European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources.
Accessibility and Clarity	European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.

3.8 Adapt your DMP: part 3



This is the third of seven 'Adapt your DMP' sections in this tour guide. After working on this chapter, you should be able to define the processing that will be done to your data during the project.

To adapt your DMP, consider the following elements and corresponding questions:

Versioning

- » How will you version your data files (and scripts) during the project?
- » Will you create and/or follow a convention for versioning your data?
- » Who will be responsible for securing that a "Masterfile" will be maintained, documented and versioned according to the project guidelines?
- » How can different versions of a data file be separated?

Interoperability

In order to be able to link your work to other research, it might be useful to build on established terminologies as well as commonly used coding and soft- and hardware wherever this is possible.

- » Which software and hardware will you use? How does this relate to other research?

If applicable:

- » Will established terminologies/ontologies (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- » Which coding is used (if any)? How does this relate to other research?

Data Quality

- » How will data quality be evaluated?
- » What data quality control measures will be used?"

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 4

Store

Contents

Main take-aways	87
4.1 Storage	88
4.2 Backup	93
4.3 Security	97
4.4 Adapt your DMP: part 4	100
Sources and further reading	102

[View the online version of this chapter](#)

Main author of this chapter

Jonas Recker, GESIS

Introduction



The data that you collect, organise, prepare, and analyse to answer your research questions, and the documentation describing it are the lifeblood of your research. Put bluntly: without data, there is no research. It is therefore essential that you take adequate measures to protect your data against accidental loss and against unauthorised manipulation.

Particularly when collecting (sensitive) personal data it is necessary to ensure that these data can only be accessed by those authorized to do so. In this chapter¹, you will learn more about measures to help you address these threats.

Main take-aways

After completing the chapter, you should be:

- » Aware of different storage solutions and their advantages and disadvantages;
- » Able to plan a storage strategy adequate to the needs of your project;
- » Able to plan a backup and disaster recovery strategy to ensure that no data loss, e.g. through human error or hardware failure, will occur during the project;
- » Able to decide when and how to protect your data against unauthorised access with strong passwords and encryption.
- » Able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt them to your own DMP.

¹ This chapter is based on information which was put together by the [UK Data Service](#) (2017), the online course [Research Data MANTRA](#) (EDINA and Data Library, University of Edinburgh, 2017) and [Essentials 4 Data Support](#) (RDNL, n.d.).

4.1 Storage

I have terabytes of videotaped interviews from a European project, dozens of pseudonymised transcripts and informed consent forms. European partners need access to the files for data analysis. What's the best storage strategy for me?

A possible storage solution

Type of data	Storage needs	Storage solution
The data which were collected are personal data.	High storage capacity for videos required;	Data are transmitted only in encrypted form. (see Security)
Extra security measures to protect it should be in place (see Security).	Remote access to videos and transcripts required;	Data for remote access is stored in cloud storage in Europe. (see Storage)
	Researchers need to work on the same files simultaneously.	Master copies of videos and transcripts are encrypted and backed up in the cloud and on portable hard disk and flash drives. (see Security)
		Backups locked away in different, secure locations. (see Backup)
		Consent forms and encryption keys are stored in a secure safe.

When choosing a suitable storage solution to fit your project's needs, a lot of questions need answering. For example:

- » How much storage space do I need?
- » Who needs access?
- » What precautions should I take to protect my data against loss?
- » Which storage solutions are suitable for personal data?

It is an important aspect of data management planning to determine what your storage needs are and select solutions accordingly. In the '[Adapt your DMP](#)' section questions that need answering are covered in more detail.

Storage solutions overview

In the following, you will find an overview of different storage solutions. Factors that play a role are, for example, data sensitivity, ease of access, file size and overall data volume. Advantages and disadvantages are detailed as well as precautions you should take when working with personal (sensitive) data. Each solution closes with recommendations on what to look out for if you decide to use the solution in question.

Portable devices

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
<ul style="list-style-type: none"> » Allow easy transport of data and files without transmitting them over the Internet. This can be especially helpful when working in the field. » Low-cost solution. 	<ul style="list-style-type: none"> » Easily lost, damaged, or stolen and may, therefore, offer an unnecessary security risk. » Not robust for long-term storage or master copies of your data and files. » Possible quality control issues due to version confusion. 	<p>Use in combination with encryption and strong password protection.</p>

Recommendations

- » Do: use for temporary, short-term storage for non-sensitive data, e.g. in the field or to transport data and files when online transmission is not possible.
- » Do: use in combination with encryption and strong password protection, especially if working with sensitive information (see ['Security'](#)).
- » Do: conduct regular checks to ensure your device is working and that files are accessible.
- » Do not: use for long-term storage or master copies of your data and files.

Cloud storage

E.g. Google Drive, OneDrive, Dropbox, a University's OwnCloud, Open Science Framework and Tresorit

Do you want to transfer personal data abroad? Read this first!

The General Data Protection Regulation (GDPR) only permits personal data to be stored within the EU, unless:

- » Participants consent to the data being stored in another country (this needs to be real consent i.e. a true choice);
- » There are adequate and equivalent levels of data protection in place (e.g. the US/EU Privacy Shield agreement).

However, researchers should assess whether they really need to store the data abroad. If data does need to be stored outside the EU then information sheets and consent forms should clearly identify this and explain the reasons why this is necessitated (See 'Informed consent').

Further guidance on sharing data outside the European Economic Area (EEA) can be found from the Information Commissioners Office.

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
<ul style="list-style-type: none"> » Automatic backups. » Often automatic version control. 	<ul style="list-style-type: none"> » Not all cloud services are secure. May not be suitable for sensitive data containing personal information about EU citizens. » Insufficient control over where the data is stored and how often it is backed up. » Free services by commercial providers (e.g. Google Drive, Dropbox) may claim rights to use content you manage and share them for their own purposes. » Data can be lost if your account is suspended or accidentally deleted, or if the provider goes out of business. 	<p>Encrypt all (sensitive) personal data before uploading it to the cloud. This is particularly important to avoid conflict with European data protection regulations if you do not know in which countries servers used for storage and backup are located (see 'Security' for more information on encryption; also see 'Protecting data').</p>

Recommendations

- » Do: use cloud services for granting shared, remote and easy access to data and other files to all involved in the project.
- » Do: Read the terms of service. Especially focus on rights to use content given to the service provider.
- » Do: Opt for European, national, or institutional cloud services which store data in Europe if possible.
 - » [B2drop](#) (EUdat, n.d.) is an example of a European cloud storage solution.
 - » [SWITCHdrive](#) (SWITCH, 2017) is a Swiss solution.
 - » [DataverseNL](#) (Data Archiving and Networked Services, 2017) is an example of a service for Dutch researchers that allows the storage and sharing of data both during and after the research period.
- » Do not: make this your only storage and backup solution.
- » Do not: use for unencrypted (sensitive) personal data.

Local storage

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
<ul style="list-style-type: none"> » Full control over files. » May be easier to protect against unauthorised access. 	<ul style="list-style-type: none"> » If data and files are stored on only one device, they are vulnerable to loss, e.g. if the device has a malfunction, is stolen or files are overwritten/erased due to human error. » Only the person who has access to the computer can access the data and files. 	<p>Protect the computer with a password and consider encrypting the hard drive.</p>

Recommendations

Using desktop computers and personal laptops as the primary way of storing and accessing data and files is only suitable for projects involving very few people (ideally: only yourself) and where data and files will not have to be moved back and forth between personal computers frequently.

If you plan to work on the data on different (local) workstations, e.g. with your laptop at home and the desktop in the office:

Do: make sure that you always work on the most current version of your files, for example with the help of versioning software or version control guidelines (see 'Data authenticity, versions and editions').

Do: make sure that the most current version is always backed up (see 'Backup').

Networked drives

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
<ul style="list-style-type: none"> » Data and files are centrally stored. » Shared access, remote access for everyone involved in the project possible. » Backups can be centrally managed and automated. 	<ul style="list-style-type: none"> » Higher security precautions are required to prevent unauthorised access and the accidental deletion or manipulation of data and files. » Access for external project partners can be difficult or impossible. » Higher cost. 	<p>Use in combination with a suitable security strategy to protect data against unauthorised access.</p>

Recommendations

- » Do: Use for distributed collaborative projects involving many people who need access to data and files
- » Do: use in combination with a suitable security strategy to protect data and files against unauthorised access (see 'Security').
- » Do: use in combination with strict versioning rules (see 'Data authenticity, versions and editions')
- » Do: think about long-term archival solutions for data that is complete and has been analysed. Valuable storage space might be released in this way.
- » Do: work with rights and permissions to ensure that not everyone has access to everything if this is not required (e.g. access to master files more restricted than access to working files).

Types of storage media

In addition to finding a storage solution that best suits the requirements of your project, you may be required to decide which media types to use for storage and backup of your data and documentation. This is of particular importance if backup and storage are not taken care of by the IT department of your university or research institute.

Optical

Example	Advantages	Disadvantages
» CD, DVD	<ul style="list-style-type: none"> » Portability » Low cost 	<ul style="list-style-type: none"> » Easily damaged, especially when handled poorly or stored under poor conditions » Easily lost » Frequent read/write errors » Not durable » Relatively small capacity

Magnetic

Example	Advantages	Disadvantages
» Hard Disk Drive (HDD)	<ul style="list-style-type: none"> » Lower cost compared with built-in Flash drives (Solid State Disks) » High storage capacity 	<ul style="list-style-type: none"> » Subject to physical degradation » Easily damaged (e.g. by magnetic fields or by physical impact)

Flash (portable)

Example	Advantages	Disadvantages
» Solid State Drive (SSD)	<ul style="list-style-type: none"> » Robustness » Relative longevity 	<ul style="list-style-type: none"> » Data hard to recover if the drive fails » Higher cost compared with magnetic Hard Disk Drives (HDD) » Smaller capacity compared with HDD

Tips for your storage strategy

The UK Data Service (2017b) recommends the following for any storage strategy:

- » **Use two types of storage media**
At least two different types of storage media should be used, e.g. Solid State Disk (SSD) and CD-ROM or Hard Disk Drive (HDD) and SSD.
- » **Replace storage media**
Replace storage media after 2-5 years.
- » **Carry out integrity checks**
Frequently carry out integrity checks to ensure that the stored data has not been corrupted. This can be done with so-called checksum tools. These allow you to detect if a file was changed in any way, intentionally or unintentionally.

How to... check the integrity of your files

We recommend that you frequently check the integrity of your files. This can be done with checksum tools such as [MD5summer](#) (n.d.) or [Checksum Checker](#) (2014). Such tools create a 'digital fingerprint' - a string of numbers - from the bit values (the ones and the zeros) of a file. Monitoring whether the fingerprint of a given file changes allows you to detect if a file was changed in any way intentionally or unintentionally.

Follow the steps in [this video](#) (UK Data Service, 2016b) to perform a checksum check for your own files.

4.2 Backup

Backups are an important instrument to ensure that data and related files can be restored in case of loss or damage. Among the most common causes of data loss are:

- » Hardware failure;
- » Software malfunction;
- » Malware or hacking;
- » Human error (research data accidentally gets deleted or overwritten or is lost in transport);
- » Theft, natural disaster or fire;
- » Degradation of storage media.

Creating a backup strategy in 10 steps

A backup strategy in one sentence would be: Make at least three backup copies of the data on at least two different types of storage media, keep storage devices in separate locations with at least one off-site, regularly check whether they work, ensure you know the process and follow it. In the list below the steps to create a backup strategy are outlined in more detail.

1. Find out whether your institution has a backup strategy

Find out whether your institution has a backup strategy. If so, backups may automatically be taken care of for any files stored on institutional servers. However, it is necessary that you check if the backup strategy in place sufficiently meets your requirements.

2. Determine what you want to back up

The three common options for backups are:

1. Full backup of the entire system and files;
2. Differential backups, where everything is recorded that was changed since the last full backup. To restore your data and/or system, you will require the last full backup and the last differential backup;
3. Incremental backups, where only changes since the last backup are recorded. To restore your data and/or system, the last full backup and the entire series of incremental backups is required.

Differential and incremental backups are also called “intelligent” backups. If only a small percentage of your data changes on a daily basis, it’s a waste of time and disk space to run a full backup every day.

3. Decide how many backups you will need and how frequently to back up

It is recommended that you make three backup copies. This will greatly minimise the risk of data loss, even in the case that one of the backups is damaged or lost. However, if storage capacity is an issue and/or if sensitive data is involved, it may be necessary to work with fewer copies.

You should clearly state in your backup strategy how often backups will be made. The frequency of backups will depend on the frequency and amount of changes to your data and documents.

4. Decide where backups will be stored

We recommend that you store at least some of the backups in (physically) different places. For example, backing up to two servers standing in the same room or building may cause you to lose both backups in case of a fire. Having an offsite copy of your backup mitigates this risk.

Backups can be made to networked drives, cloud storage, and to local or portable devices (see [‘Storage’](#)). What works best for your project depends on the amount of data that needs to be backed up, the required frequency of backups, the level of automation, and the sensitivity of the data.

5. Determine how much storage capacity will be needed

Estimate which amount of data and documentation you will collect and create in your project. Then determine the corresponding approximate amount of storage capacity needed for backups. If your institution has an IT department, they will be able to help you with this.

6. Determine if there are tools you could use to automate backup

Automating backups can help to ensure that backups are created at the correct time and that they are saved to the correct location, reducing the risk of human errors. Both Microsoft and Apple operating systems have software to support automatic backups. Cloud storage solutions too often have a backup functionality. However, make sure to check frequently that functional backups were indeed created.

- » OS X
Have a look at the [video tutorial on creating backups for your Mac](#) using Time Machine (UK Data Service, 2016b).
- » Windows 10
Windows 10 includes two different backup programs:
 - » File history
The File History tool automatically saves multiple versions of a given file, so you can “go back in time” and restore a file before it was changed or deleted. That’s useful for files that change frequently.
 - » Windows Backup and Restore
The Backup and Restore tool creates a single backup of the latest version of your files on a schedule.

Of course, you would still need an off-site backup as well.

7. Determine how long backups will be kept and how they will be destroyed

It is generally recommended that you do not overwrite one backup with another. However, if you have to back up large amounts of data frequently it may not be feasible to retain all backups for the entire duration of the project.

If sensitive data is involved, make sure that any deleted data are truly gone and cannot be recovered in any way. For suitable procedures, see [‘Security’](#).

8. Determine how personal data will be protected

Make sure that backups of data containing sensitive information are protected against unauthorised access in the same manner as the original files. For suitable measures, see the chapter on [Security](#).

9. Devise a disaster recovery plan

A disaster recovery plan defines the steps to take if a data loss occurred and thus helps you to restore data as quickly as possible. The plan should also assign responsibilities for data recovery tasks and list persons (or functions) to contact when a data loss occurs.

To ensure that data recovery will run as smoothly as possible in the event of an actual data loss, make sure to regularly test whether restoring lost files from your backups is actually possible.

10. Assign responsibilities

Never assume that someone will take care of backups and data recovery. Assign responsibilities for making manual backups, for checking those automatic backups actually happened, for testing data recovery, and for restoring any lost data.

Determine how to check the integrity of backed-up files

Errors can happen when backups are written or copied. We recommend that you frequently check the integrity of your backed up files. This can be done with so-called checksum tools such as [MD5summer](#) or [Checksum Checker](#).

The UKDS compares checksums to [digital fingerprints](#). Available tools create such a fingerprint with the help of an algorithm that computes the fingerprint - a string of numbers - from the bit values (the ones and the zeros) of a file. Monitoring whether the fingerprint of a given file changes allows you to detect if a file was changed in any way intentionally or unintentionally.

Video tutorial on using MD5summer: <https://www.youtube.com/watch?v=VcBfkB6N7-k>

Case studies

In the following, two scenarios will be used to illustrate the importance of backups and to highlight some of the things that are important to consider when planning a backup strategy. After reading through the scenarios, take a few minutes to think about what could have been done to prevent data loss. Afterwards, you can open the tabs to see our diagnosis.

Lost backpack

On a night out after work, a friend's backpack was lost containing literally all of their data and documents for their Master's thesis. A fairly recent copy of the thesis text is backed up in DropBox, but the only two copies of the data - video-recordings and transcripts of interviews with primary school teachers in rural areas of Ireland - were on the laptop (transcripts and sequences from the videos) and the hard drive (original, unanonymised videos and backed-up files from the laptop). Both were lost with the backpack.

Analysis: Measures employed to protect data and participants

- » The thesis text was backed up to the external hard drive and to the cloud;
- » Transcripts and video sequences from the laptop were backed up to the external hard drive but not to DropBox because they contained sensitive information;
- » No backup of the video footage existed. The entire footage was on the external hard drive in unencrypted form.

Measures that could have reduced the negative effects of data loss

1. Keep backups in different locations

One thing the scenario illustrates is that when it comes to backup, never put all your eggs in the same basket. No matter how many backups you have - if all of them are in the same place, the risk to lose everything is considerable. For storage, consider the advantages and disadvantages of different storage solutions and storage media (see [Storage](#)).

A rule of thumb is to keep three backups, at least one of them in a different location from the others, on different types of storage media. However, sometimes considerations of privacy or storage capacity will require you to deviate from this recommendation.

2. Use encryption to protect research participants' privacy

In the scenario, the lost hard drive contained personal data of participants in the research. The loss, therefore, compromises the privacy of the involved individuals. Whenever personal data is stored and processed for research, backup measures have to be linked with data protection measures. Personal data should be encrypted and anonymized as quickly and comprehensively as the research objective permits. You should also create only as many copies of this data as absolutely required. Note that this may involve diverging from the "three copies" rule mentioned above.

Master copy gets overwritten

A group of researchers collaboratively works on quantitative survey data. They use a shared working space on a networked drive where a master copy of the data and a working copy are stored. Two backups exist, stored separately from the working and master files: one copy on an external hard drive and one in the university's own Cloud system.

A new researcher enters the project. Who is not aware of the way files are named and organised and accidentally works on the master copy of the data. In this process, a number of variables get overwritten when the new team member recodes variable values and forgets to save them into a new variable. Fortunately, two backups exist.

The researchers know that sometimes copies can get changed due to write or transmission errors, so they decide to check with a checksum tool if the two copies are identical. They discover that the checksums for the two files are not identical. This means that either one or both of the files were altered in some way.

Analysis: Measures employed to protect the data

- » The master copy is kept as a separate file from working files;
- » Two backup copies on different media and in different locations exist;
- » No frequent integrity checks of the backups were made and no additional protection for the master copy of the data was in place to prevent it from being overwritten.

Measures that could have reduced the negative effects of data loss

1. Versioning and file naming rules

Errors such as accidentally overwriting a file can always happen, but they are less likely to occur if clear rules for versioning and file naming are in place and if folders are clearly labeled. Such policies and guidelines help to avoid confusion about what files contain and where they should be saved. See '[Data authenticity, versions and editions](#)'.

2. Restricting access to important files

As mentioned above, human error is one of the most common causes of data loss. Therefore, consider restricting the access to important files, for example with the help of passwords or by using systems with read and access rights management. By giving fewer people access to important files, the risk of data loss caused by human error can be minimised.

3. Creating three backup copies rather than only two

If three copies rather than only two had been created, this would have increased the chances of identifying the unaltered copy: if two out of three copies are identical, this suggests that these are unharmed. This would have saved the project laborious work of trying to identify the correct copy.

4. Checking the integrity of files

Errors can happen when backups are written or copied. These can sometimes make a copy entirely unusable, but sometimes they are small enough to go unnoticed initially but then cause problems further down the line. This could lead to you losing access to the data entirely - for example because a software can suddenly no longer render the files - or it can cause the data to contain errors, thus impacting the results of your research negatively. Learn more about integrity checks in this [video about performing a checksum check](#) for your files (UK Data Service, 2016a).

4.3 Security

To prevent unauthorised access and possible changes to your data, data security measures are in order. Such measures, on the one hand, serve to protect personal data and confidential information and on the other hand offer protection against unauthorised manipulation or erasure of files (intentional or unintentional).

Data security can be considerably increased with the help of technical measures. However, these must be accompanied by organisational measures in the form of policies and guidelines.

Measures

In the video linked below, several measures that directly contribute to data security are detailed: limiting access with passwords, encrypting data and disposing of data that you no longer need securely. These measures are exemplified and supplemented by other measures in the boxes below (LSI Storage, 2009).

<https://www.youtube.com/watch?v=Ylkg7-JOYX8>

Passwords

To protect your data files, you should use passwords to lock the computer systems used to access these data files. The University of Edinburgh (2017) [has compiled some guidance](#) on how to choose a strong password. In general, they should be long (15 characters or more). A very useful way to choose strong passwords is to make them up of four randomly chosen and altered words, e.g. C.r3ctHorseBatteryStaple.

Edward Snowden (LastWeekTonight, 2015) [advises us](#) to shift our attention away from passwords to pass phrases which are unlikely to be in a dictionary, e.g. MyMotherM\$kesTheB*stCakes. This way of thinking does not only make passwords stronger, but also a lot easier to remember.

The video (Alexanderlehmann, 2015) below explains why pass phrases are hard to crack. It is in German, but you can turn on English subtitles.

<https://www.youtube.com/watch?v=jtFc6B5lmIM>

Password security

Besides choosing strong passwords, make sure to store and transmit them securely so they cannot be stolen:

- » Do: store passwords in a sealed envelope in a secure place (e.g. a safe);
- » Do: use secure password management tools. Remembering all of your passwords can be a challenge. Password management tools are one possibility of dealing with this problem. Examples are [KeePassX](#) (2017) and [Lastpass](#) (2017);
- » Do not: write passwords down and leave them lying about openly (e.g. in your desk drawer);
- » Do not: enter passwords in untrustworthy environments such as open wifi or internet cafés.

Encryption

Encryption is the process of encoding digital information in such a way that only authorised parties can view it. It is especially useful when you are transmitting personal or confidential data.

When you encrypt a file, the information it contains is “translated” into meaningless code. To translate this code back into meaningful information a key is required. Attacks with ransomware such as the [Locky virus](#) (“Locky”, 2017) have demonstrated that recovering information from encrypted files without the key is nearly impossible. It is therefore extremely important that you do not lose the key to decrypt your files.

- » **Do:** encrypt confidential data, especially before transmitting it online, uploading it to the cloud, or transporting it on portable devices. When working in a team, make sure that the key can be accessed by everyone who needs to access it (but only those people).
- » **Do:** ensure that you do not lose the key to decrypt your files, e.g. by keeping it in a sealed envelope in a secure location such as a safe room

Encryption software

The UK Data Service (2017c) has [compiled information on encryption](#) and offers short video tutorials demonstrating the use of different software tools to encrypt data.

Commonly used encryption software includes:

- » [BitLocker](#) (2017)
Standard on selected editions of Windows. For the encryption of disk volumes and USB devices.
- » [FileVault2](#) (Apple Inc, 2017)
Standard on Apple Macs. For full disc encryption.
- » [PGP \(Pretty Good Privacy\)](#) (Raicea, 2017)
There are commercial programmes (e.g. by Symantec (Symantec Corporation, 2017)) and free/open programmes (e.g. Gnu Privacy Guard (GnuPG, 2017)) available.
- » [VeraCrypt](#) (n.d.)
Multi-platform encryption software (Windows, Mac and Linux). For full disk and container encryption.
- » [Axcrypt](#) (n.d.)
Open source file-level encryption tool with free and commercial versions available for Windows and MacOS.
- » [SafeHouse](#) (2012)
Free and commercial software versions available for Windows. Encrypts files, folders and drives.

Physical, network and computer security

To prevent your data from being manipulated or stolen, sufficient security measures to block any unwanted access to rooms and buildings or computers and networks where they are held should be in place.

- » Do: log and/or control access to physical sites where sensitive information is stored, e.g. with the help of key cards.
- » Do: use strong passwords and encryption (see above).
- » Do: use up-to-date virus scanners and firewalls.
- » Do: ensure that systems used to access data are continually updated (e.g. security updates for the operating system).

The UK Data Service (2017d) has a [list of further important security measures](#).

Secure disposal

Used Phones Are Full of Previous Owners' Data: Researchers bought 20 used smartphones in four cities, and recovered thousands of photos, texts, and emails | Wadell, 2016.

Managing your data also means thinking about how to securely dispose of confidential information. Merely hitting the “delete” button on your computer or mobile device is not enough. In fact, even formatting the hard drive or doing a factory reset can leave (portions of) confidential information in place.

There are two options for secure disposal of confidential data:

- » **The physical destruction of the storage medium** (e.g. shredding of discs)
- » **The use of software for secure erasing**
There are [various software options available](#) (UK Data Service, 2017e) that can securely delete files from hard drives. For example, [AxCrypt](#) (n.d.), [Eraser](#) (2017) and [WipeFile](#) (2014) are free open source file and folder shredding utilities.

The UK Data service (2017e) [points out](#) that solid-state hard disks (SSD) and USB flash drives (memory sticks) use a different technology than hard drives. Therefore, the techniques for securely erasing files are also different. The use of manufacturer-specific software is recommended. Note, though, that especially for solid state drives and USB flash drives only physical destruction is a 100% guarantee that the data cannot be recovered.

Contact the IT department and the administration of your university or institute to find out about regulations and procedures for secure destruction of confidential data.

Organisational aspects

Data security partly depends on technological and physical protection measures. However, these measures alone are not sufficient and will not adequately protect your data if you do not also address the “human factor”. This is particularly important if working collaboratively in a bigger and/or distributed team.

Protection against security breaches depends on the establishment and communication of clear rules and guidelines. Here are some points to consider when planning your data management that focus on the human/organisational dimension of data security:

Do: Invest time to draw up policies and concrete guidelines/checklists for all topics discussed in this chapter, especially:

- » Passwords: minimum requirements for password strength; management/secure storage of passwords.
- » Encryption: what types of data are encrypted for which purposes using which tools?
- » Secure data transmission and transport.
- » Secure data disposal.

Do: Restrict access to sensitive data:

Most likely, not everyone on the team needs access to all files. Determine who needs access to which types of data and handle access restrictions, e.g. with the help of passwords. In addition, create a routine to ensure you adapt authorisations in case someone leaves the team.

Do: Create awareness and keep communication going:

Errors often happen due to a lacking awareness of potential issues or threats. For example, does everyone on the team know which data is considered sensitive and why? Is everyone aware of potential risks posed by transmitting unencrypted data via email? Make sure that everyone on the team is adequately involved in discussions of data security issues and measures in place.

4.4 Adapt your DMP: part 4



This is the fourth of seven 'Adapt your DMP' sections in this tour guide. For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

After working on this chapter, you should be able to define your storage and backup strategy for your data and metadata. To adapt your DMP, consider the following elements and corresponding questions:

Short-term storage strategy

Type of data

Are you collecting personal data or do your data in any other way require special protection?

Who needs access

- » Is it necessary to have remote access to the data? Are you e.g. transmitting data from the field?
- » How important is fast access?
- » Is simultaneous and synchronised access by several people required?

Storage capacity

- » How much data are you going to generate and how much storage capacity will you need, including backups?
- » Which media types will you use and how often will you replace them?

Storage period

For how long is storage required?

Data security

- » How will you protect your data? (passwords, encryption, physical, network and computer security measures?)
- » How will the data be disposed of (if need be)?

Backup procedures

- » How many backups will you make and where will these be stored?
- » How will the integrity of backups and disaster recovery be tested?

Budget

What is the associated cost of storing and backing up data?

Long-term storage strategy

Thinking about storage as part of data management planning also entails considering what will happen to your data and files after the project ends. Where and for how long will data be retained? While the recommended route is to [archive and publish your data](#) at the end of the project by handing it over to a data repository, for some data this may not be possible or desired. Maybe no consent was given for sharing, or publication would infringe intellectual property rights of third parties.

To ensure that your data remains accessible even after the end of the project consider the following questions:

Storage period

For how long after the project is the data and the documentation to be kept? 10 years after you last published an article based on the data is commonly considered the minimum period for retention unless legal or ethical issues require shorter or longer retention periods (e.g. see the [funder retention requirements of UK funding council policies](#) listed on a Libguide by the University of Southampton (2017)).

Storage location

- » Where will the data and documentation be kept after your project ends?
- » Can you or your employer guarantee sufficiently secure storage and backup for the data for the envisioned retention period without losing access?

File formats

- » Are you certain that your data and files are stored in a format for which there will still be suitable software available to access and process the stored information in ten years?
- » Which file formats will you use to minimise the risk that current software can no longer read your data files? See '[File formats](#)' for further information.

Budget

What is the associated cost of storing and backing up data and documentation after the project has ended?

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 5

Protect

Contents

Main take-aways	104
5.1 Ethics and data protection	105
5.2 Ethical review process.....	107
5.3 Processing personal data	113
5.3.1 Diversity in data protection.....	120
5.4 Informed consent.....	125
5.5 Anonymisation.....	133
5.6 Copyright	140
5.6.1 Diversity in copyright	143
5.7 Adapt your DMP: part 5	153
Sources and further reading	154

[View the online version of this chapter](#)

Main authors of this chapter

Scott Summers, UK Data Archive

Libby Bishop, UK Data Archive

Introduction



This part of the tour guide focuses on key legal and ethical considerations in creating shareable data.

We begin by clarifying the different legal requirements of the European Union Member States, and the impact of the General Data Protection Regulation (GDPR) on research data management. Subsequently, we will show you how sharing personal data can often be accomplished by using a combination of obtaining informed consent, data anonymisation and regulating data access. The supporting role of ethical review in managing your legal and ethical obligations is also highlighted in this chapter.

DISCLAIMER: This chapter is based on European and country-specific laws and codes of research ethics. Any guidance and advice within this module do not constitute legal advice. Professional legal advice should be sought where you are unsure of the legal requirements placed upon you by law when conducting your research.

Main take-aways

After completing this chapter you should:

- » Be aware of your legal and ethical obligations towards participants and be informed of the different legal requirements of EU Member States;
- » Understand how protecting your data properly protects you against violating laws and promises made to participants;
- » Understand the impact of the General Data Protection Regulation ([GDPR](#); European Union, 2016a);
- » Understand how a combination of informed consent, anonymisation and access controls allows you to create shareable personal data;
- » Be able to define what elements should be integrated into a consent form;
- » Be able to apply anonymisation techniques to your data;
- » Be able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt your own DMP.

5.1 Ethics and data protection

When collecting, using and sharing research data, ethical considerations and legal obligations guide the way.

Ethics are an integral part of a research project, from the conceptual stage of the research proposal to the end of a research project. Within the EU the RESPECT project has drawn up [professional and ethical guidelines](#) (Institute for Employment Studies, 2004) for conducting socio-economic research. The RESPECT Code of Practice is based on three main guidelines:

1. Upholding scientific standards

Researchers should always seek to take account of all the relevant evidence and present their research without omission, misrepresentation or deception.

This means in practice that researchers should ensure that when they are formulating their research questions, designing surveys, questionnaires or interviews they do not predetermine or prejudice the outcome through their choice of questions or actions.

2. Compliance with the law

Researchers need to ensure that they are aware of all the relevant national and international laws that may affect their research project. With collaborative projects which cross legal borders, this may involve various laws. Ones of particular relevance (and to be aware of) will be in regards to data protection and intellectual property. These will be discussed in more depth in this chapter.

3. Avoidance of social and personal harm

Researchers should aim to avoid or minimise social harm to groups or individuals when conducting their research projects. This means that the research project should be designed responsibly and consider participants throughout. For example, participation in the research project should be voluntary and on the basis of fully informed consent.

Depending on the type of data you collect you will have to deal with different laws. Whereas Intellectual Property legislation applies to all data, the collection of personal data has its own laws to adhere to. Importantly, since 25 May 2018, the [General Data Protection Regulation](#) (GDPR; European Union, 2016a) applies to any EU researcher or researcher in the European Economic Area (EEA) who collects personal data about a citizen of any country, anywhere in the world, as well as any researcher worldwide who collects personal data on EU citizens.

Archiving and publishing personal data

Recap: What are personal data?

If you collect research data that enables you to identify a person, then this is classified as personal data. Within the General Data Protection Regulation ([GDPR](#), European Union, 2016a) personal data is defined as any information relating to an identified or identifiable natural person known as 'a data subject'. It is further specified that an identifiable natural person is someone who can be identified, either directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Personal data can include a variety of information, such as names, address, phone number and IP addresses.

The GDPR applies only to the data of living persons. Data which do not count as personal data do not fall under data protection legislation, though there may still be ethical reasons for protecting this information.

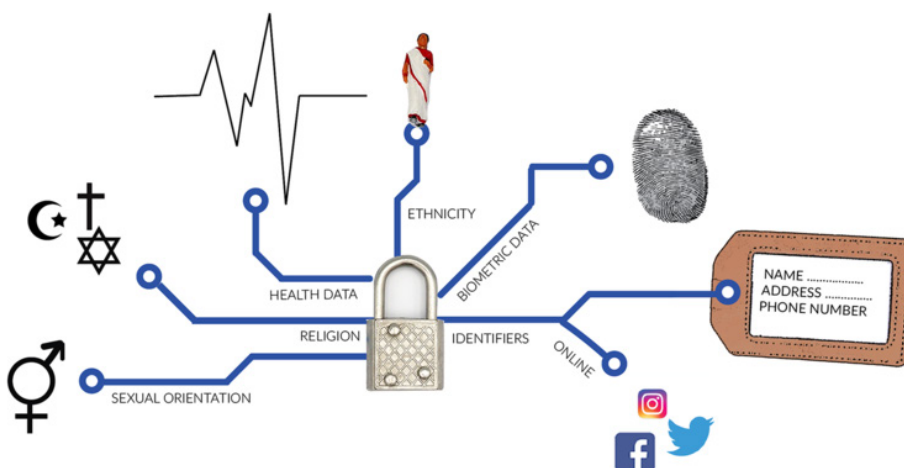
Sensitive personal data

Certain personal data are considered particularly sensitive and thus require specific protection when they reveal information that may create important risks for the fundamental rights and freedoms of the involved individual. Examples of sensitive personal data include data revealing religious affiliation, sexual orientation, or racial or ethnic origin. Within the GDPR the following categories are defined as ‘special categories of personal data’:

- » Racial or ethnic origin;
- » Political opinions;
- » Religious or philosophical beliefs;
- » Trade union membership;
- » Genetic data;
- » Biometric data;
- » Data concerning health;
- » Data concerning a natural person’s sex life or sexual orientation.

There are other data which may contain sensitive information which do not fall under the special categories of personal data but should still be treated as such, including, for example, confidential business data and secret data concerning state security.

Many research funders and journals expect or require data sharing (i.e., data to be made available in a data repository). Especially for (sensitive) personal data, there may be a perceived tension between data sharing and data protection. In the coming paragraphs, we will show how a combination of gaining consent, anonymising data, gaining clarity over who owns the copyright to your data and controlling access to data can enable the ethical and legal sharing of data.



First, we will get you started on the topic of ethical review. Starting with an ethical self-assessment will help you identify the key ethical and legal issues in your study beforehand, which will maximise your data’s value whilst protecting your participants.

5.2 Ethical review process

Ethical review is about helping you as a researcher to think through the ethical issues surrounding your research. The principles of good research practice encourage you to consider the wider consequences of your research and engage with the interests of your participants.

Ethics review by a Research Ethics Committee (REC) is typically required when (sensitive) personal data are being collected. The role of a REC is to protect the safety, rights, and well-being of research participants and to promote ethically sound research. Among other duties, this involves ensuring that research complies with national and international data protection laws regarding the use of personal information collected in research.

Ethical self-assessment

Regardless whether there is a formal requirement, we recommend you to perform an ethical self-assessment. The type of questions which are generally to be answered in an ethical review is shown in the illustration below. These questions are derived from the [Ethical guidelines for research](#) by the Norwegian National Research Ethics Committees (n.d.)

» **Question 1: The project's aim and method**

Could the project's aims and methods come into conflict with commonly recognised values? Could carrying out the project involve risk of injury to people, animals, or nature to an extent that should not be neglected? If so, are the persons involved aware of the risk?

» **Question 2: Research involving identifiable persons**

Does your research involve personal data collection and/or processing? If so, will informed consent be obtained from the participants? Will personal information be sufficiently anonymised in order to ensure adequate privacy protection?

» **Question 3: Whistle-blowing**

If a project employee develops serious doubts regarding ethical aspects of the project, will he or she be allowed to present his or her worries to an independent consultative body? Is this option made known in advance?

Ethical review in H2020

[Since FP7](#) (the European Union's Research and Innovation funding programme for 2007-2013; European Commission, 2013)) and its successor [Horizon 2020](#) (running from 2014-2020; European Commission, n.d.) the EU has started to require ethical review. We have used the [H2020 ethical guidelines](#) (European Commission, n.d.) to get you acquainted with the steps which may be taken in an H2020 ethical review process. Do note that step four will be very rare.

1. Ethical self-assessment

The first step when applying for funding under the H2020 scheme is for the applicants to [perform an ethics self-assessment](#) (European Commission (2016), pages 16 and 17) to submit with their research proposal. This entails completing an ethical checklist about how you will protect your participant's personal data and involves considering questions around how the data will be collected and stored securely and safely, how the data will be retained, and whether any of the data will be transferred to any non-EU countries.

2. Ethical screening

The ethical screening process takes place during the scientific evaluation of the proposal or soon after it is considered for funding. It takes into account the ethical self-assessment conducted by the researcher. If an ethical issue is identified the ethical aspects of the proposals objectives, methodology, and potential impact will be considered by ethical experts.

3. Ethical assessment

In limited cases, an ethical assessment may need to be undertaken, which involves an in-depth analysis of the ethical issues of the proposal. The conclusions of the ethical screening are also taken into account. This typically happens in cases where there will be severe intervention on humans.

4. Ethics check

During the ethical screening or ethical assessment, the experts identify the projects that need an Ethics Check to be executed during the course of the research project. In case of substantial breach of ethical principles, research integrity or relevant legislation, the Commission can afterwards carry out an Ethics Audit. The checks and audits can result in an amendment of the grant agreement.

European diversity in ethical review

Ethical guidelines for research involving people are often issued by professional bodies, host institutions, and funding organisations. Therefore, these rules and guidelines will differ from country-to-country as well as from research funder to research funder. This wide variation in requirements for ethical review across countries is challenging, especially for multi-national research projects. Usually, when working in more than one country, the strictest regulations typically apply. It is also good practice to engage with any local regimes, where possible.

Below we list examples of local diversity in ethical review. Also see the [online version of this guide](#).

Croatia

According to the [Act on Scientific Activity and Higher Education](#), the Croatian Parliament appoints the [Committee for Ethics in Science and Higher Education](#) which shall promote ethical principles and values in science and higher education, business and public relations, and in the application of advanced technologies and environmental protection. The Committee adopted the [Code of Ethics](#) determining principles of ethics in higher education, the publication of results, relations among scientists, teachers and other participants in the process of science and teaching, procedures and activities related to market competition, as well as relations to the public and the media.

By the same law, higher education institutions, scientific research institutes and other scientific research organisations may, in accordance with the statute, establish their own ethics committees and adopt their own codes of ethics, which must comply with the Code of Ethics of the Committee for Ethics. For example, Faculty of Humanities and Social Sciences [Department of Sociology](#) and [Department of Psychology](#) have formed their own committees and published guidelines and procedures on how and why researchers and students should submit their research proposal for ethical review. In their work, they follow national and international professional codes of ethics in addition to general ethical principles. In the case of research on children, applicants are advised to familiarize themselves with the [Code of Ethics for Research on Children](#) (the new version is soon to be published).

The Croatian Science Foundation, the major funder of basic, applied and developmental research adopted the [Code of Ethics of the Croatian Science Foundation](#). This Code of Ethics contains a set of principles in the area of scientific integrity and scientific ethics that serve as guidelines for professional activities and public actions of all Foundation's employees, members of the Foundation's bodies and boards, evaluators and beneficiaries of the Foundation's funds, as well as other researchers whose work is connected with the Foundation's activities. The Code is based on

the European Code of Conduct for Research Integrity, which establishes best scientific practice on the principles of scientific integrity, guiding the researchers in their encounters with practical, ethical and intellectual challenges, including reliability, honesty, respect and accountability.

Czech Republic

There exist several ethical codes and standards that apply to empirical social research in the Czech Republic. The general ethical principles for research are introduced in the [Ethical Framework for Research](#), the set of recommendations (not obligations) approved by the Czech government.

Large research organizations in the Czech Republic have their own ethical codes and committees, for example [Charles University's Code of Ethics](#) or [Code of Ethics for Researchers of the Czech Academy of Sciences](#).

There are professional associations in different fields of social science research and humanities, e.g. [Czech Sociological Association](#) or [Czech-Moravian Psychological Society](#). While the psychological society has defined field-specific ethical rules and maintains ethical committee, the Czech Sociological Association is lacking both specific ethical rules and ethical committee and it is often a problem to get any official expert opinion on ethical issues from this organization.

Data collection for research purposes, even the academic ones, is usually conducted by credible commercial research agencies. Such agencies are members of international and national professional organizations (SIMAR, resp. ESOMAR) and adhere to recognized standards, codes of ethics and other rules (ISO 20252 etc.). Individual researchers are often members of professional organizations (EFAMRO, WAPOR)

Germany

Ethical review of Social Science research

Research ethics has received increased attention over the past years. Besides regulations like Federal and State Data Protection Laws (Datenschutzgesetze, see the website of the [Federal Commissioner for Data Protection and Freedom of Information](#) for an overview), there are voluntary ethical statutes of organizations like the German Association of Sociology ([DGS](#), German only), the German Association for Political Science ([DVPW](#), German only), or the German Association for Psychology ([DGP](#), German only).

The German Ethics Council ([Deutscher Ethikrat](#)) publishes policy-relevant guidance on all areas of research ethics (mostly German, English Abstracts available). The German Data Forum ([RatSWD](#)) has published recommendations and teaching material on research ethics (German only). It also hosts a(n) (incomplete) list of local ethics commissions. The incompleteness (as of December 2019) is due to the fact that many individual universities are just now beginning to set up general Ethical Review Boards that handle all types of human subjects research. Finally, the [Data Ethics Committee](#) has also published a broad expertise for ethical conduct concerning the handling of research data in general under the auspices of the Federal Data Protection Commissioner (German only).

Intellectual Property Rights and data

Concerning data archiving, creators of data and documentation are treated as owning the Intellectual Property Rights of the research data. When archiving, they transfer the non-exclusive rights of use and rights of reproduction to the archive on the basis of an archive agreement. The non-exclusive rights of use comprise the right to pass data and documentation to a third-party and the right to change the format of digital objects (data files, tables, etc.) for the purpose of long-term preservation (e.g. migrating files to the latest formats). All incoming datasets are checked for possible ethical and data protection issues (voluntary consent, direct personal references, etc.).

The agency Euraxess offers an overview of [German intellectual property rights](#).

North Macedonia

Article 14 of the Law on scientific and research work (2008) establishes a national level Ethics' board, 9 member body, 6 of which are proposed/appointed by the Intra-university conference and 3 by the Macedonian academy of sciences and arts. The main function of the board is "monitoring and evaluation of ethical principles and values in scientific work, protection of human integrity in scientific research, and ethics in professional relations among those performing scientific research".

The Board adopted an Ethical code that addresses ethical principles in: scientific work, in the publication of the results of scientific work, in the relations among researchers, in the procedures and activities related to competition, and in the relations with the public and public media.

Universities are not obliged, but they can adopt their own Ethical codes. Biggest universities have done this.

Research disciplines in which ethical behavior is of utmost importance (medicine, psychology etc.) have adopted their own Ethics codes.

The Law on personal data protection (2005) also addresses important issues that, although not explicitly, apply to ethics in scientific research.

Norway

The [Norwegian National Research Ethics Committees](#) (n.d.) are independent agencies for questions regarding research ethics and investigation of misconduct, within all subject areas. All the committees provide ethical guidelines on research ethics within the different subject areas.

The Regional Committees for Medical and Health Research Ethics (REK) must give prior approval for medical and health research projects and general research biobanks. REK may also grant exemption from the duty of confidentiality for health information used for non-medical research.

For other subjects (science, technology, social sciences, law, humanities and research on human remains) there are advisory bodies (NENT, NESH and The National Committee for Research Ethics on Human Remains) for research ethics in its subject areas which provide advice and recommendations for specific projects submitted to the committees. Obtaining advice prior to a research project is not mandatory, but researchers are encouraged to contact the committee if the project is considered to present challenges in terms of research ethics.

NSD – The Norwegian Centre for Research Data AS, as a [Data Protection Service](#), offers research institutions an agreement to assess the processing of personal data in research projects in accordance with data protection legislation.

Serbia

Within the Ministry of Education, Science and Technological Development of the Republic of Serbia, the National Council for Scientific and Technological Development functions as the highest decision-making body. On February 21, 2018, the Council adopted a [Code of Conduct for Scientific Work](#), which the Ethics Committee for Science takes care of. Apart from this central body, each accredited university, faculty, and institutes have an obligation to produce its own document that will regulate ethics in the scientific research of a particular institution.

Slovenia

The three main Universities in Slovenia adopted their Codes of Ethics: [University of Ljubljana](#), [University of Maribor](#) and [University of Primorska](#) (only in Slovenian language). Committees for Ethics in Research are established on the level of faculties. For instance, Committee for Ethics in Research at the Faculty of Social Sciences (CER FSS) examines the applications for ethical assessment of research tasks and projects undertaken at the Faculty of Social Sciences involving research work that interferes with the privacy of people or engages in research involving people. The Committee discusses the applications of teachers, researchers and research associates employed at FSS, and students upon the proposal of a mentor. The Committee gives opinion on proposals of research projects that involve research work with people, using methods of humanities and social sciences.

On the national level there are National Medical Ethics Committee, Administration for Food Safety, Veterinary Sector and Plant Protection, which discuss the ethical issues of the relevant field of research, and if needed the researcher shall obtain the opinion from the relevant body or organization.

Sweden

The [Act concerning the Ethical Review of Research Involving Humans](#) (2006:460) was implemented with the purpose to protect the individual person and ensure respect for human dignity in research. It includes provisions with a requirement for ethical review of research involving living and deceased persons or biological material from humans.

In Sweden, if you are going to process personal data, and you work with research at a Swedish university or authority, you need a Data Protection Official for Research to help you ascertain that you follow the GDPR and Swedish legislation.

Research may only be approved if the risks it may entail to study participants in regard to health, safety, and personal integrity are outweighed by the scientific value. Research cannot be approved if the expected results can be reached in a way that presents fewer risks to study participants. Research may only be approved if it is to be conducted by, or under the supervision of, a researcher who possesses the necessary scientific competence.

The Act includes provisions on information to study participants. Researchers are required to inform study participants of the overall plan for the research, the purpose of the research, the methods that will be used, the consequences and risks that the research may entail, the identity of the research principal, and the facts that participation is voluntary and that participants can withdraw their participation at any time. An ethical review is always required, even if study participants have given their expressed consent to the use and handling of their data.

Switzerland

In Switzerland, the legislation makes it mandatory for a research project to be evaluated by a cantonal commission when it falls within the scope of the [Swiss Federal Act on Research on Human Beings](#) (Human Research Act, HRA). The HRA's scope of application is "any project for which biological material is collected from a person or personal data related to his or her health are collected in order to respond to a scientific problem or to reuse biological material or health-related data for research purposes" (Art. 6). All researchers working on subjects related to the diseases, structure and functioning of the human body, or at least those working with personal data related to these subjects should, therefore, consult their cantonal commission to determine whether or not they are subject to the HRA. If so, they must submit their project for evaluation before any data is collected.

For research projects that do not fall within the scope of the HRA, there is no legal obligation to be evaluated by an

ethics committee. However, in some universities there are committees through which it may be mandatory to go through, depending on the type of subject being studied. Other universities offer such commissions as a service for research that requires ethics validation in order to meet the increasing demands of funders, publishers, fields, disciplines, etc.

United Kingdom

In the UK, some form of ethical review is required for most research involving human participants, personal (sensitive) data or controversial methodologies (e.g., covert research). Funders, universities, journals, or other bodies may make these requirements. The major funder of social research in the UK, the [ESRC](#), requires reviews to be completed prior to the start of research (but not when submitting a proposal) (ESRC, 2017a). On their website [guidance is offered](#) (ESRC, 2017a).

Most institutions offer a graduated review system of review, ranging from a self-assessment checklist to a light-touch review for most student and low-risk projects, to comprehensive review at the institutional level. [This flowchart](#) (Economic and Social Research Council, n.d.) helps in deciding what type of ethical review your project needs. Furthermore, [the case studies](#) (ESRC, 2017b) may help you in gaining a picture of the ethical dilemma's which may arise during your own research project.

Expert tips

TIP 1. Educate your REC



RECs may be informed and supportive of efforts to share data. However, there is great variation, and some oppose data sharing, fearing (mistakenly) that sharing data violates participants' confidentiality. As a researcher, you may need to ensure that your REC is fully informed on these subjects. At a minimum, REC members should know that:

- » Many research funders and journals expect or require data publication (i.e., data to be made available in an archive or repository);
- » Consent forms should allow for participants to opt in or opt out of data sharing, whilst also protecting their confidentiality (see ['Informed consent'](#));
- » Data protection laws only apply to personal data, but they do not apply to anonymised data;
- » Identifiable information may be exempt from data sharing;
- » A combination of gaining consent, anonymising data and controlling access to data can enable the ethical and legal sharing of data; even sensitive data can be shared if suitable procedures and precautions are taken, as is done at major data repositories.

Tip 2. Finding RECs

Find the REC at your own institution or have a look at The European Network of Research Ethics Committees - [EURECis](#) (EUREC, n.d.) - which brings together already existing national Research Ethics Committees (RECs) associations, networks or comparable initiatives on the European level.

5.3 Processing personal data

Since 25 May 2018, the [General Data Protection Regulation](#) (GDPR, European Union, 2016a) applies to any EU researcher or researcher in the European Economic Area (EEA) who collects personal data and any researcher worldwide who collects personal data on EU citizens. The GDPR applies only to the data of living persons. Data which do not count as personal data do not fall under data protection legislation, though there may still be ethical reasons for protecting this information.

The [GDPR](#) (General Data Protection Regulation, Chapter 2, Article 5) prescribes that you should adhere to the following six principles when processing personal data:

I. Process lawfully, fairly and transparently

The participant is informed of what will be done with the data and data processing should be done accordingly.

II. Keep to the original purpose

Data should be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

III. Minimise data size

Personal data that are collected should be adequate, relevant and limited to what is necessary.

IV. Uphold accuracy

Personal data should be accurate and, where necessary kept up to date. Every reasonable step must be taken to ensure that personal data that are inaccurate are erased or rectified without delay.

V. Remove data which are not used

Personal data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

VI. Ensure data integrity and confidentiality

Personal data are processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.

The research exemption

The GDPR contains an exemption which entails that some of the principles above are slightly different when you collect and process personal data for research purposes. This is called the 'research exemption'.

Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subjected to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner | [General Data Protection Regulation, Article 89](#).

In practice, this means that Principle II. and V. are less strict. Further processing of personal data for the purposes of archiving, scientific or historical research purposes and statistical purposes is not considered to be incompatible with the initial purposes of data collection, even when this purpose was not expressly mentioned earlier.

Also, personal data may be stored for longer periods for such purposes. In all cases, appropriate technical and organisational measures should be taken to safeguard the rights and freedoms of the participants in your research, such as data minimisation and pseudonymisation.

Legal Basis

Personal data can only be processed when there is a valid legal basis to do so. The GDPR recognises six bases (grounds):

- » consent of the data subject
- » necessary for the performance of a contract
- » legal obligation placed upon the data controller
- » necessary to protect the vital interests of the data subject
- » carried out in the public interest or in the exercise of official authority (public task)
- » legitimate interest pursued by the data controller

In research, the three most applicable bases for processing personal data are consent, public interest (public task) or legitimate interest. For each research project, if personal data will be collected and processed, the most appropriate legal basis needs to be decided and recorded (and should not be changed at a later date). The UK Data Service has [published examples](#) of where a legal basis may be applied in research.

GDPR in practice

When you start a research project that involves collecting information from people, for example via a survey or interviews, then consider whether or not you will collect personal data. If not, then data protection legislation does not apply. If you will collect personal data, then:

- » determine who will be the data controller (possibly your institution)
- » decide which legal basis will apply
- » if collaborative partners need access to personal data, then make sure agreements are in place
- » consider whether a Data Protection Impact Assessment is needed (see details on this in the GDPR Questions and Answers below)
- » communicate to research participants how personal data collected about them will be used, stored, processed, transferred, who the data controller is (with their contact details), the legal ground and purpose of the processing, the period of retention and their rights; this can be done via an information sheet or a webpage (e.g. privacy notice)
- » consider where to store personal data securely
- » minimise the personal data to collect and pseudonymise where possible

GDPR Questions and Answers

Below are some questions and answers about how to implement the GDPR requirements in practice in a research project, resulting from [2019 CESSDA Webinar](#).

Q: I am a postdoc researcher doing a qualitative study, interviewing women about abusive relationships. I will use pseudonyms for each woman interviewed. Respondents may still be identifiable from the story they tell. Does this constitute personal information? If so, which legal ground should I use for this research?

A: Yes, this would constitute personal information. In this case the legal ground could be consent, which should be sought from the women participating in the study. Another aspect to keep in mind here is data collected which would allow identification of other people who may not have been asked for consent, for example partners carrying out the abuse. So you may also be processing personal data from people who have not been asked for consent. In that case, the processing ground could be public interest and the argument would be that the research has value for society. If the project allows, such partners could be made aware of the processing of their data, if this poses no risks to the participating women.

Q: I am doing an online poll survey, using Qualtrics, asking 5000 people across Europe for which political party they voted in the recent European elections, also recording their ethnicity and other demographic information. Does this qualify as processing special categories data? If so, how do I gain explicit consent for collecting this information?

A: A first consideration would be how much identifying/personal information is collected during the survey, alongside the political view and ethnicity. This helps to decide whether this classifies as special categories data. If no data is collected that allows identification of the respondents, then the GDPR will not apply. If identifying information is collected, then this qualifies as special categories data and therefore explicit consent would be needed. One way to achieve this would be through double consent, whereby consent for processing personal data collected would be asked at the beginning and the end of the questionnaire.

Qualtrics is a USA based company and thanks to negotiations by various European survey institutions, Qualtrics now only processes collected survey data in the EU for EU-based surveys. This means that Qualtrics can be used as a tool for surveys that need to comply with the GDPR.

Q: What are the GDPR rules when using administrative or register data that contain personal information?

A: If consent is not collected from the individuals when the administrative or register data are collected, then the most common legal basis for further use is public task. If consent can be sought, that would be preferable.

Q: The GDPR indicates strongly that a consent form should be easy and clear, yet I have to provide so much extra information to my interviewees now. How do I do this?

A: The best way to provide this information to participants is through an information leaflet and a consent form. You can provide the information in a written leaflet. If you are interviewing people you can explain the leaflet content also face-to-face to make sure it is people understand the content.

Q: If a researcher brings an electronic device across the border to a third country, sends an email or publishes personal data on the web, does this constitute as a data transfer?

A: An email containing personal data sent from Europe to someone in a non-European country would indeed constitute a data transfer. An electronic device containing personal data carried across the border to a third country would constitute a data transfer if the personal data will be passed on to another person. If personal data are published on the web, it depends on whether the data are stored and who can access them. If it is openly published it could be considered a data transfer.

Q: What are the data protection implications for international partnerships and research projects when non-EU countries are involved?

A: If personal data are going to be handled/processed as part of the partnership research activities within the EU, then the GDPR would apply. One solution would be that the European-based partners require their non-EU partners to have appropriate privacy/data protection measures in place and that consent is given by all subjects, irrespective of whether they are based in Europe or not. That may not always be easy or possible. However, solutions can be found such as data anonymisation, data encryption, using secure servers and partners can learn from each other. Good practice is also for all users and purposes of use of the personal data to be recorded.

Q: Does the GDPR apply to personal data, collected outside the European Economic Area (EEA) and transferred to the EEA for analysis?

A: Yes, it would, because it would be classified as personal data once stored within the EEA.

Q: Are there examples of research where using consent as legal basis for processing personal data would not be suitable?

A: Covert research is an example where consent would not be an appropriate processing ground, as asking for consent would have a negative outcome for the research. In covert research, public task would likely be the best ground. It is still important that the research adheres to ethical principles, and the researcher is open about the process used in publications.

Q: How can we comply with the GDPR when studying populations that are easily identified, for example surveys of candidates running in a general election or surveys of the members of a scientific association?

A: First, you need a legal basis for the processing of personal data. The most common legal basis for this scenario may be consent. If you gain consent from the people studied you can give information about the risk of being identified in published outcomes of the survey and ask consent on that basis. If the legal basis for processing personal data is public task, you should give information about the study to the population to make sure that they can manage their rights according to the GDPR.

Q: How is the 'right to be forgotten' applied in research settings?

A: The right to be forgotten applies in research, but is not an absolute right. Best practice is to inform participants about this right as clearly as possible and explain what it means and what it may not mean. For example, if data have been published in which people are identifiable, for example a paper containing a quote for which permission was given. Then if a participant wants to be forgotten, it would be very difficult to retract the paper. So be clear to participants about what they can do with this right and up to which point they can withdraw from research and request to be forgotten.

Q: Is a Data Protection Impact Assessment (DPIA) only required in scientific research for sensitive data concerning vulnerable subjects?

A: A DPIA is required for data processing that is likely to result in a high risk to the rights and freedoms of individuals. In practice this means if at least two of these criteria apply (examples can be found in the Data Protection Working Party 248 guidelines):

- » evaluation or scoring
- » automated-decision making with legal or similar significant effect
- » systematic monitoring
- » sensitive data
- » data processed on a large scale
- » datasets that have been matched or combined

- » data concerning vulnerable data subjects
- » innovative use or applying technological or organisational solutions
- » data transfer across borders outside the European Union
- » when the processing in itself prevents data subjects from exercising a right or using a service or a contract.

At the same time, a DPIA is a good learning tool. For a research project that involves the collection of personal data, a joint session of the researcher with a legal person and a technical person is very useful to establish best practices for data protection. This helps to understand context and helps to define common problems, solutions and risk mitigation measures.

Q: How are Data Protection Impact Assessments (DPIAs) being implemented across different institutions, for research?

A: If research is done as a collaboration of more than one institution, with shared responsibilities, one DPIA done by one of the institutions should be enough, and the other partner institutions should apply that same DPIA. Problems might arise when research involves institutions that are implementing a DPIA in different countries, whereby policies or requirements may vary across those countries, such as for data security, ownership of the data, different understandings on gaining consent and which legal basis to use for processing personal data.

Q: How should researchers deal properly with the GDPR in the context of open data?

A: For personal data, the open access motto “as open as possible, as closed as necessary” is important. A political or societal drive for open access and open science does not mean that individual rights granted by legislation can be overruled. Therefore, for personal data, ‘as closed as necessary’ is the key.

Q: What is the applicability of ‘legitimate interests’ in research using Artificial Intelligence (AI)?

A: The use of AI is a specific form of using personal data, and legitimate interest could be a legal basis for AI. More important is the framework provided by guidelines and recommendations of the High-Level Interest Group on AI: [Ethics guidelines for trustworthy AI](#) and [Policy and investment regulation for trustworthy AI](#).

Q: When a US entity is a processor of pseudonymised EU citizen data and the key to re-identify the subjects exists only in the EU, so that the US entity cannot re-identify the subjects, does the GDPR apply to the US entity? Is the US entity required to sign a contract if requested by the EU entity?

A: If the US entity has no access to the key, then the data would in theory be classified as anonymous data. If the key would ever be released or the US entity would gain access to it, then the data would be defined as pseudonymised data or personal data. The organisation would need to decide whether signing a processor agreement would be best, considering the risks they wish to take.

Q: In research projects that plan to use data collected from social media platforms, how can researchers reconcile the right to privacy vs. the publicly available data?

A: Gaining consent would be the best approach when using social media data. So even for social media data in the public domain, researchers should ask the people whose social media content they mine for their consent when possible. In some cases public task could be used as legal basis.

Q: Are European countries converging or diverging in their choice of legal basis for processing personal data in research across Europe, specifically when considering whether consent or public task would be used in research?

A: The UK strongly encourages the use of public task as legal ground in research, whereas many other European countries favour consent. The UK view may pose a risk for participants’ rights. We will be

able to evaluate in future how this has evolved. For the German case one can rather see a diverging trend since the Federal government left things open to be defined by the 16 Federal States. They took the chance and eight have now introduced the definition of “anonymous data” as formerly used in the German Data Protection Act. But they all see consent as a major basis for research.

Q: Should data repositories and data archives be considered as data processors or data controllers? Is archiving research data from a project part of the original processing for the research, or does it constitute a separate, further processing?

A: In most instances it is likely that data archives would be considered as data processors. However, some data archives may also be involved in undertaking research for the projects, which could lead to them being a joint controller with the research institute.

Different data archives in different countries may take a different view. Some would consider all data collections as potentially personal data and treat them as such. Only when data are considered to be fully anonymous would the GDPR no longer apply. Other archives take a two-tiered approach having certain procedures for anonymous data and other for personal data. An archive can archive personal data if there is a legal basis to do so. Liaising with the research project team is important.

Secure disposal

Used Phones Are Full of Previous Owners' Data: Researchers bought 20 used smartphones in four cities, and recovered thousands of photos, texts, and emails | Wadell, 2016.

Managing your data also means thinking about how to securely dispose of confidential information. Merely hitting the “delete” button on your computer or mobile device is not enough. In fact, even formatting the hard drive or doing a factory reset can leave (portions of) confidential information in place.

There are two options for secure disposal of confidential data:

- » **The physical destruction of the storage medium** (e.g. shredding of discs)
- » **The use of software for secure erasing**
There are [various software options available](#) (UK Data Service, 2017e) that can securely delete files from hard drives. For example, [AxCrypt](#) (n.d.), [Eraser](#) (2017) and [WipeFile](#) (2014) are free open source file and folder shredding utilities.

The UK Data service (2017e) [points out](#) that solid-state hard disks (SSD) and USB flash drives (memory sticks) use a different technology than hard drives. Therefore, the techniques for securely erasing files are also different. The use of manufacturer-specific software is recommended. Note, though, that especially for solid state drives and USB flash drives only physical destruction is a 100% guarantee that the data cannot be recovered.

Contact the IT department and the administration of your university or institute to find out about regulations and procedures for secure destruction of confidential data.

Organisational aspects

Data security partly depends on technological and physical protection measures. However, these measures alone are not sufficient and will not adequately protect your data if you do not also address the “human factor”. This is particularly important if working collaboratively in a bigger and/or distributed team.

Protection against security breaches depends on the establishment and communication of clear rules and guidelines. Here are some points to consider when planning your data management that focus on the human/organisational dimension of data security:

Do: Invest time to draw up policies and concrete guidelines/checklists for all topics discussed in this chapter, especially:

- » Passwords: minimum requirements for password strength; management/secure storage of passwords.
- » Encryption: what types of data are encrypted for which purposes using which tools?
- » Secure data transmission and transport.
- » Secure data disposal.

Do: Restrict access to sensitive data:

Most likely, not everyone on the team needs access to all files. Determine who needs access to which types of data and handle access restrictions, e.g. with the help of passwords. In addition, create a routine to ensure you adapt authorisations in case someone leaves the team.

Do: Create awareness and keep communication going:

Errors often happen due to a lacking awareness of potential issues or threats. For example, does everyone on the team know which data is considered sensitive and why? Is everyone aware of potential risks posed by transmitting unencrypted data via email? Make sure that everyone on the team is adequately involved in discussions of data security issues and measures in place.

5.3.1 Diversity in data protection

It is one of the key responsibilities of researchers and the Project Principal Investigators to familiarise themselves with the local laws, rules and ethical requirements for their projects.

When research crosses legal and jurisdictional boundaries researchers should always seek to apply the requirements of the legislation that has the most stringent requirements of the whole project. Where this is unclear, you should obtain advice from your institute, ethical committees or qualified legal professionals.

Since 25 May 2018, the [General Data Protection Regulation](#) (GDPR; European Union, 2016a) applies to any researcher who collects data on EU citizens. One of its key aims is to harmonise laws across the EU regarding data protection legislation.

In addition to the GDPR, each EU Member State has rules on data protection and legislation that you have to familiarise yourself with if you collect personal data. Because of this, some Member States have more restrictive data protection legislation than others.

Key national legislation affecting data protection is presented in the tables below. For up-to-date information, see also the [online version of this guide](#).

Croatia

Personal data protection is a constitutional right, in the framework of human rights and fundamental freedoms. "The safety and secrecy of personal data shall be guaranteed for everyone." ([The Constitution of the Republic of Croatia](#), Article 37).

[The Act on Implementation of General Data Protection Regulation](#) ("Official Gazette" No. 42/18) was enacted on 25th May 2018 to ensure full implementation of the GDPR in Croatia. National legislation contains no provisions regulating the use of personal data in scientific research.

The [Croatian Personal Data Protection Agency](#) is the only independent public supervisory authority in the Republic of Croatia within the meaning of the provision of Article 51 of the General Data Protection Regulation. The Agency can be contacted by researchers for consultation services about the use of personal data in their research.

Finland

Informing participants

Under the Finnish Data Protection Act it is a requirement that when personal data are collected or processed for research that participants are informed about the purpose of the research and what will happen to their contribution [[Personal Data Act \(523/1999\)](#), 1999].

Potential participants must have enough information to be able to make an informed choice on whether to partake in the research or not. Before collecting personal data in Finland, researchers must fill in the 'Description of the scientific research data file'. Ethical review boards usually require this file, and research participants have the right to see it, should they wish to do so. In cases where the personal data are drawn from registers (and no consent has been asked from the participants), the description of the scientific research data file must also be sent to the Office of the Data Protection Ombudsman.

More information and advice can be found at the [Finnish Social Science Data Archive](#) (2017a). You can also contact the [Office of the Data Protection Ombudsman](#) (n.d.) directly.

Germany

Data protection in Germany is governed by the GDPR, the Federal as well the State Data Protection Acts.

There is no centralized authority for research ethics and data protection due to the federal nature of Germany. The [Federal Commissioner for Data Protection and Freedom of Information](#) is part of the [Data Protection Conference](#) (webpage in German only).

The German Data Forum ([RatSWD](#)) has published recommendations and teaching material on research ethics and data protection (webpage in German only). The [Federal Data Protection Act](#) was adapted to the GDPR in 2017.

Greece

Data protection in Greece is governed by:

-Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), known as GDPR

-Law 4624/2019 with the title "Hellenic Data Protection Authority (HDP), measures for implementing Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data, and transposition of Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 and other provisions"

Law 2472/1997 which has been repealed, except for the provisions referred to expressly in Article 84 of Law 4624/2019-Law 3471/2006 on the protection of personal data and privacy in the electronic telecommunications sector

Law 3471/2006 with respect to the electronic communications sector which incorporates into the Greek law European Directive 58/2002.

The above regulatory framework sets out the obligations of those who process personal data and the respective rights of those to whom the data processing relates. The same Law also provides for the establishment of the Hellenic Data Protection Authority (HDP) and its powers and competencies. The Hellenic Data Protection Authority (HDP) is a constitutionally consolidated independent Authority which incorporates into the Greek law. relevant EU legislation provisions.

All relevant, official, publicly available information can be found [on the HDP website](#).

Netherlands

Research data should be stored permanently as far as possible insofar as scientists participate in research by or at the institution that has adopted the [Netherlands Code of Conduct for Research Integrity](#) (Association of Universities in the Netherlands, 2018) or discloses their research in its name, research findings and research data should be made public subsequent to completion of the research to the extent possible. Simultaneously, the institutions have the obligation to ensure permanent storage as far as possible. Upon receiving a grant from the Netherlands Organization for Scientific Research ([NWO](#)) it is important to address this in the [Data Management Plan](#) (DMP).

Norway

Project notification

In Norway, if you are going to process personal data and you work at one of the institutions that have an agreement Norwegian Centre for Research Data (NSD) as their Data Protection Services for Research then you must notify NSD about the research project. If your institution does not have an agreement with NSD, you must either notify your institution's own Data Protection Official (if they have one) or the Norwegian Data Protection Authority. A notification is not required only if the research project registers anonymous information only. However, you should note that you will still need to notify the NSD if you will be processing personal data during the project, even if the research project will publish anonymous data.

If you are a researcher employed at an institution outside Norway different rules apply: if the data controller (i.e. the responsible institution) is established in an EEA country, it is sufficient to submit a notification of the project to the relevant authorities in the country concerned. If the data controller is located in a country outside the EEA, the notification must be submitted in Norway by a Norwegian institution that undertakes the role of the data controller's representative.

Further information and advice can be sought from the [NSD](#) (n.d.) directly.

North Macedonia

In this area, research institutions are obliged to respect the general provisions of the Law on personal data protection (2005) (Official gazette of the R. Macedonia No.7/05, 103/08, 124/08, 124/10, 135/11, 43/14 and 153/15). The new law, which implements the GDPR directive of the EU, is in preparation. Anyway, research institutions have developed their own practices for the protection of personal information during the research process.

Serbia

The Government of the Republic of Serbia adopted the Law on Personal Data Protection of the Republic of Serbia, in November 2018 and its implementation Act began on August 21, 2019. The Law on Personal Data Protection puts personal data at the very top of the protection priorities and gives citizens the functional capacity to manage their privacy much better and more transparently. In essence, the Law is the final product of a general civic and political initiative that has launched a long-standing "hard-fought" process to achieve legal frameworks in which each individual would have greater protection of his or her privacy and in which institutions and companies would be given much clearer rules and procedures by which they could to process and use personal information.

The special rule applies to data processing for a purpose archiving in the public interest, scientific or historical research, and statistical purposes, as well as when it comes to the right of access to information, or in general relationship between the right to protection of personal data and freedom of expression.

Slovenia

In Slovenia, the Personal Data Protection Act ([Slovene, English](#)) is still not adopted to General Data Protection Regulation. Researchers can find some guidelines on this topic at the Information Commissioner office. For more, see [Publications and Guidelines of the Slovenian Information Commissioner](#).

Sweden

The Swedish Ethical Review Authority is a recently restructured authority under the Ministry of Education and Research, for the protection of humans in research, research on biological material and sensitive personal data.

On their [website](#), researchers can find information on the legal requirements that must be complied with in order for the research to be legal. The ethical rules that apply to research are based on international conventions, founded on principles for research ethics. Swedish research is covered by international law and conventions, as well as national legislation. The rules are there to make sure that individuals are not harmed or subjected to unnecessary risks when personal data and people are used for research.

The General Data Protection Regulation, GDPR, and complementary legislation includes all usage of personal information. Personal information can, according to GDPR, only be used for specific, explicitly stated, and legitimate purposes. However, even when all these criteria are met, personal information for research purposes also requires informed consent.

There are exceptions in case there is a rule that conflicts with other Swedish constitutional law, for example conflicts with The Freedom of the Press Act (SFS 1949:105), or The Fundamental Law on Freedom of Expression (SFS 1991:1469). Treatment of personal information for artistic or journalistic purposes is excepted, as well as private registers.

The [SND website](#) provides further information (in Swedish).

Switzerland

Data protection in Switzerland is both regulated at the federal and the cantonal level. At the federal level, it follows the [Federal Act on Data Protection](#) (FADP) (The Federal Council, 2014) and the [Ordinance to the Federal Act on Data Protection](#) (OFADP) (The Federal Council, 2012). Besides the FADP, each of the 26 cantons has their own cantonal data protection act. Universities are regulated by cantonal laws.

The FADP is currently under revision and should align with the GDPR. See [Finsterwald](#) (2016) for more practical information.

UK

In the UK, there is the Freedom of Information Act and a common-law tort of breach of confidence.

Freedom of Information Act

Researchers who work at a publically funded research institute or university in the UK are subject to the Freedom of Information (FOI) Act 2000. This Act provides members of the public with a right to access information held by UK public sector organisations (e.g. publically funded research institutes and universities). This means that a member of the public may make a request for access to a researcher's research data.

There have been various examples of research data being requested through the FOI Act. For example, climate change researchers at the University of East Anglia had two such requests made in early 2007. The university initially refused to release data, however after one of the requesters drafted a letter to the ICO alleging that the university was in violation of the FOI Act the university released the requested research data (Booth, 2009).

An [FOI request](#) (GOV.UK, n.d.) can come in many forms, but for it to be valid, it must come in a written form, such as an email, letter or fax. An FOI request can also come from anyone, meaning that the requester does not have to have been a participant in the research project. The information needs to be provided unless an exemption or exception allows the researcher not to disclose the information. Researchers must respond within 20 working days of receiving the request and should seek assistance from their university/research institute before disclosing any

information. This is particularly important where the FOI request requests access to data which is not that of the requester but is defined as 'personal data' under the GDPR of another 'data subject'.

Researchers working on European projects need to be aware that they will need to comply with the UK FOI Act if there is a UK public research institute or university involved in their research project.

Further [guidance on FOI](#) (ICO, n.d.a) can be sought from your research institute/university or the [UK's Information Commissioner's Office](#) (ICO, n.d.b).

Breach of confidence

In the UK, there is a common-law tort of breach of confidence. A duty of confidence arises when confidential information comes to the knowledge of a person in circumstances where it would be unfair if it were then to be disclosed to others.

Disclosure of information subject to a duty of confidentiality would constitute a breach of the duty. The duty of confidentiality is not absolute and is not protected by legal privilege, and exceptions occur. For example, where the participant has consented to the information being used in specific ways, for agreed purposes, and by certain people or where a judge requires disclosure.

This applies to information not already in the public domain. If the consent form promises confidentiality, disclosing information unlawful may constitute a breach of confidence.

5.4 Informed consent

The following statement has been adapted from an actual consent form: “**Any information I give will be used for research only and will not be used for any other purpose**”. Consider the implications for data sharing for any data generated using this consent statement. Do you have any suggestions for alternative wording or other changes?

Some thoughts on this statement

Some comments/reflections:

1. It is tempting to use such wording as a way of reassuring participants that their data will not be misused, but this may be overly restrictive.
2. Perhaps the data—appropriately anonymised—could be equally useful for teaching, for example.
3. In general, think very carefully about any wording that restricts – forever – uses of the data. If what you are trying to do is to build trust with participants, telling them how their data can be safely used in diverse ways is a better approach!

Informed consent is the process by which a researcher discloses appropriate information about the research so that a participant may make a voluntary, informed choice to accept or refuse to cooperate.

Normally informed consent is given before the start of the research. Gaining informed consent is crucial to meeting your legal and ethical obligations towards participants whilst simultaneously enhancing the value of your research data.

To obtain informed consent, researchers should:

- » Inform participants about the purpose of the research;
- » Discuss what will happen to their contribution (including the future archiving and sharing of their data);
- » Indicate the steps that will be taken to safeguard their anonymity and confidentiality;
- » Outline their right to withdraw from the research.

Consent needs to be freely given, informed, unambiguous, specific and by a clear affirmative action that signifies agreement to the processing of personal data.

Examples of consent forms

Sample consent forms are available in the online version of this guide:

<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/5.-Protect/Informed-consent>

Information sheets

Information sheets play an important role in gaining a participant’s informed consent to take part in a research project. They help provide participants with the background information which is necessary to make an informed decision about whether to take part in the research project.

A good information sheet discusses the following topics:

- » The purpose of the research;
- » What is involved in participating in the research;

- » The benefits and risks of participating in the research;
- » Details of the research, e.g. the funding source, sponsoring institution, name of project, contact details for researchers and how to file a complaint;
- » The procedures for withdrawing from the research project;
- » The planned usage of the data during the research, dissemination, storage, publishing and archiving of the data;
- » The strategies for assuring ethical use of the data;
- » The procedures for safeguarding personal information, maintaining confidentiality and anonymising data, particularly in relation to data archiving, sharing and reuse.

Examples of information sheets

Sample information sheets are available in the [online version of this guide](#).

Gaining informed consent for data archiving and sharing

Gaining informed consent for data sharing is seen as ‘one more small step’ to gaining consent from participants to partake in your research project. As a researcher, you will already be acutely aware of the need to fully inform your participants about:

- » What taking part in your research project will involve;
- » How you will disseminate information from the project through publications or presentations;
- » The impact taking part may have on them.

By adding the discussion of data sharing and archiving you permit the participant to make an informed decision. This empowers them and puts them in charge of choosing whether they wish for their contribution to your research project – and their data – to be available for use in future research projects.

Granular consent

The best way to achieve informed consent for data sharing is to identify and explain the possible future uses of their data and offer the participant the option to consent on a granular level. For example, in a qualitative study, this may involve allowing the participant to consent to data sharing of the anonymised transcripts, the non-anonymised audio recordings and the photographs.

Below, an example of what granular consent for data sharing could look like on a consent form is detailed.

The interviews will be archived at and disseminated so other researchers can reuse this information for research and learning purposes:

I agree to the non-anonymised audio recording of my interview being archived and disseminated for reuse yes/no

I agree to the anonymised transcript of my interview being archived and disseminated for reuse yes/no

I agree to any photographs of me taken during interview being archived and disseminated for reuse yes/no

Approaches to informed consent

Consent can be gained from participants in written or oral form, one-off or continuously throughout the research project, retrospectively or not at all. The form of consent sought will depend on the project. In the accordion below the details and considerations of all three are stated.

Written or verbal consent

Choice	Advantages	Disadvantages
Written consent	<ul style="list-style-type: none"> » More solid legal ground, e.g. participant has agreed to disclose confidential info; » Often required by Ethics Committees; » Offers more protection for researchers (as they have written documentation of consent). 	<ul style="list-style-type: none"> » Not possible in some cases: infirm, illegal activities; » May scare people from participating (or have them think that they cannot withdraw their consent).
Verbal consent	<ul style="list-style-type: none"> » Best if recorded. 	<ul style="list-style-type: none"> » Can be difficult to make all issues clear verbally; » Possibly greater risks for researchers (in regards to adequately proving participant consent).

Written consent is typically seen as the preferred form of the two options, where possible because the participant can be given detailed written information which can then be explained to them to ensure they fully understand what they are consenting to.

One-off or process consent

Choice	Advantages	Disadvantages
<p>One-off consent is where the participant is asked to consent to taking part in the research project only once.</p> <p>This would typically be at the beginning of the project before the data is collected, but could also happen at the end of the first interview.</p>	<ul style="list-style-type: none"> » Simple; » Least hassle to participants. 	<ul style="list-style-type: none"> » Research outputs not known in advance; » Participants will not know all info they will contribute.
<p>Process consent is where the participant's consent is requested continuously throughout the research project.</p> <p>For example, this may be before the first interview then after each subsequent follow up interview.</p>	<ul style="list-style-type: none"> » Ensures 'active' consent 	<ul style="list-style-type: none"> » May not get all consent needed before losing contact; » Repetitive, can annoy participants.

Retrospective consent

In cases where consent was not sought at the point of research, it may be possible to gain retrospective consent from the participants for the depositing of the data in a repository. However, if participants cannot be traced, depositing the data in a repository will need to be assessed on a case-by-case basis to identify whether it is appropriate to share it. This assessment will need to consider various factors such as the nature of the project, the consent sought, the questions asked and the anonymisation levels utilised.

On the UK Data Service (2017b) website you can [read a case study](#) on gaining retrospective consent from a 30-month research project concerning the 2001 foot and mouth disease epidemic in the UK. A standing panel of 54 local people from North Cumbria produced more than 3,000 weekly diaries about the impact of the crisis and the process of regeneration.

Expert tips



TIP 1. Documenting consent

The GDPR requires that researchers document consent if consent is the legal basis for processing personal data. An obvious way to do this is by using written consent forms. If that is not possible in the research, then verbal consent discussions and agreements can be audio-recorded if the participants agree. Or the consent process and wording used can be written out in detail.

TIP 2. Delivering informed consent in the best way possible

Researchers should consider the participant's needs, understanding and the best way to gain informed consent. This may, for example, require pictures to be used instead of lots of text – to make it clear and easy for the participant to understand – or for the consent form to be translated into the participant's native language.

TIP 3. Consent for surveys

For surveys, where personal identifiers such as people's names are not collected or are easily removed from the data file, written consent is often not gathered. Instead, the information sheet given to participants or the survey introduction state that consent for the data being used for specified purposes is implied from participating in the survey, with a clause stating that an individual's responses would not be used in any way that would allow their identification. It is therefore vital that the information sheet provides details about plans for data sharing. This information should include where the data will be deposited and the potential future uses of the data.

TIP 4. Research without consent

There are circumstances where no form of consent can be obtained for research, e.g. when the researcher collects the information from sources other than the persons themselves or when the data were already collected for another purpose. These situations are exceptional and will need case-by-case review and clear arguments before that research can be conducted. In jurisdictions which have Research Ethics Committees (REC), researchers will need to satisfy the requirements of these research ethics review boards. E.g. in Norway, NSD or the Norwegian Data Protection Authority would need to be informed and permit the research. A [Notification form](#) (NSD, n.d.c) listing the reasons why gaining informed consent isn't possible should be handed in.

European diversity in informed consent

Apart from being good scientific practice, in some countries gaining informed consent is mandated by law. Below, a consent requirement comparison for several European countries is given.

An up-to-date version of this comparison is also available in the [online version of this guide](#).

Croatia

National legislation

The [Act on Implementation of General Data Protection Regulation](#) does not regulate consent for scientific research, only for conditions applicable to child's consent in relation to information society services, and to the processing of genetic and biometric data.

Consent required to conduct research?

Depends on circumstances/In accordance with GDPR.

Verbal or written consent?

Either is permitted/In accordance with GDPR (i.e., the controller must be able to prove the consent).

One-off or process consent?

No/in accordance with GDPR (easy with exceptions for academics - but not one-off).

Czech Republic

National legislation

[Act No. 110/2019](#) Coll. on personal data processing

Consent required to conduct research?

Depends on circumstances/In accordance with GDPR.

Verbal or written consent?

Either is permitted/In accordance with GDPR (i.e., the controller must be able to prove the consent).

One-off or process consent?

No/in accordance with GDPR (easy with exceptions for academics - but not one-off).

Germany

National legislation

Data Protection Act(s); e.g. [Federal Data Protection Act](#)

Consent required to conduct research?

If personal data is collected, stored, or processed.

Verbal or written consent?

Written, but exceptions permissible (e.g. in cases where written consent would hamper the research or where other important reasons prevent obtaining written consent).

One-off or process consent?

Not defined.

Netherlands

National legislation

GDPR [National implementation act](#) (Dutch only).

Consent required to conduct research?

Depends on circumstances/In accordance with GDPR.

Verbal or written consent?

Either is permitted/In accordance with GDPR (i.e., the controller must be able to prove the consent).

One-off or process consent?

No/in accordance with GDPR (easy with exceptions for academics - but not one-off).

North Macedonia

National legislation

Law on personal data protection (art. 6, 8 and 2)

Consent required to conduct research?

Yes, article 6 of the Law on personal data protection.

Verbal or written consent?

Not defined. Both verbal and written consent possible.

One-off or process consent?

Not defined.

Norway

National legislation

[Personal Data Act](#) (Norwegian only). Information in English ([2019](#)).

Consent required to conduct research?

As a main rule yes, but there are exceptions for research.

Verbal or written consent?

Both allowed.

One-off or process consent?

Not defined, but in practice often process consent in long-term research.

Serbia

National legislation

Law on Personal Data Protection of the Republic of Serbia.

Consent required to conduct research?

Yes, but there are exceptions for research.

Verbal or written consent?

Both allowed (but preferably written).

One-off or process consent?

Not defined.

Slovenia

National legislation

New Data Protection Act is being processed.

Sweden

National legislation

[Ethical Review](#) (in Swedish only).

Consent required to conduct research?

Yes, if research is carried out on humans, biological material, or sensitive personal data.

Verbal or written consent?

Written consent is required.

One-off or process consent?

One-off.

Switzerland

National legislation

[Federal Data Protection Act](#).

Consent required to conduct research?

The Federal Data Protection Act requires consent for any processing of personal data.

Verbal or written consent?

When a research project falls within the scope of the Swiss Federal Act on Research on Human Beings (Human Research Act, HRA) and its ordinance (Human research Ordinance, HRO), consent must be explicit and written. The persons concerned (e.g. research participants) must receive comprehensible oral and written information on (HRA,

Art. 16):

- » the nature, purpose, and duration of, and procedure for, the research project;
- » the foreseeable risks and burdens;
- » the expected benefits of the research project, in particular for themselves or for other people;
- » the measures taken to protect the personal data collected; and
- » their rights,

and on (HRO, Art. 8):

- » the effort involved and the obligations arising from participation;
- » their right to withhold or to revoke their consent without giving reasons;
- » the consequences of revoking consent to further use of the biological material and personal data collected up to this point;
- » their right to receive information at any time in response to further questions;
- » their right to be informed of results concerning their health, and their right to forgo such information or to designate a person who is to take this decision for them;
- » the measures envisaged to cover any damage arising from the research project, including the procedure in the event of a claim;
- » the main sources of financing for the research project; and
- » other points relevant to their decision on participation.

Procedures and modalities of consent, however, vary according to the risks entailed by the data collection methods. More precisely, projects considered “low risk” (category A), such as those based on observation and questionnaires, benefit from a lighter informed consent approach:

- » the information may be given to the participants in successive stages and in a form other than text (HRO, Art. 8);
- » consent may be given and documented in a form other than written form (oral consent), provided that the research project is carried out with adults capable of discernment (Art. 9);
- » the possibility of using personal health-related data even after the revocation of consent, provided that the data are anonymised (Art. 10).

Despite these exceptions, researchers are never exempt from the obligation to inform participants in advance of the conditions and objectives of the project (Art. 8), and to guarantee the protection of personal data collected and/or used (Art. 5).

United Kingdom

National legislation

There is no legislative requirement for consent to be sought from participants. However, many funders, RECs, and ethics guidance bodies require it.

Consent required to conduct research?

No.

Verbal or written consent?

Either is permitted.

One-off or process consent?

Either is permitted.

5.5 Anonymisation

I am collecting data on asylum seekers' and refugees' experiences of forced labour. These participants can be considered 'doubly vulnerable'. We want to share these data. How should we protect our participant's anonymity?

A possible approach

Consider:

- » not recording any official identifying data (e.g. Home Office numbers);
- » letting participants choose their own pseudonyms (which should not be disclosive in any way);
- » password-protecting interviewee contact details;
- » not connecting pseudonyms to these password protected interviewee contact details.

Read more about the ethical considerations of this real-life project at the site of the [Economic and Social Research Council](#) (ESRC, 2017c).

The best way to protect your participant's privacy may be not to collect certain identifiable information at all. The second best is anonymisation which allows data to be shared whilst protecting participant's personal information. Anonymisation should be considered in the context of the whole project and how it can be utilised alongside, informed consent and access controls. For example, if a participant consents to their data being shared then the use of anonymisation may not be required.

Personal data can be disclosed through two categories of identifiers.

- » **Direct identifiers** are ones like the participant's name, address, or telephone numbers that specifically identify them;
- » **Indirect identifiers** are ones that when they are placed with other information could also reveal an individual, for example, by cross-referencing occupation, salary, age, and location.

Anonymisation versus pseudonimisation

Pseudonymisation and anonymisation are two distinct terms which fall under different categories in the [General Data Protection Regulation](#) (GDPR; European Union, 2016a). Whereas anonymisation irreversibly destroys any way of identifying the data subject, in theory, pseudonymisation allows to re-identify the data subject with additional information.

The GDPR defines pseudonimisation as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information". To pseudonymise a dataset "the additional information must be kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". Directly identifying data is held separately and securely from processed data to ensure non-attribution.

Anonymous data is data that cannot identify individuals in the dataset in any way. Neither directly through name or social security number, indirectly through background variables, nor through a list of names or through an encryption formula and code/scrambling key.

Anonymisation methods

When anonymising, data identifiers need to be removed, generalised, aggregated or distorted. Below, best practices for anonymising quantitative and qualitative data are given.

Quantitative data

Best practices for anonymising quantitative data:

- » This may involve removing or aggregating variables or reducing the precision or detailed textual meaning of a variable;
- » Aggregate or reduce the precision of a variable such as age or place of residence. As a general rule, report the lowest level of geo-referencing that will not potentially breach respondent confidentiality;
- » Generalise the meaning of a detailed text variable by replacing potentially disclosive free-text responses with more general text;
- » Restrict the upper or lower ranges of a continuous variable to hide outliers if the values for certain individuals are unusual or atypical within the wider group researched.

Qualitative data

Best practices for anonymising qualitative data:

- » Using pseudonyms or generic descriptors to edit identifying information, rather than blanking-out that information;
- » Plan anonymisation at the time of transcription or initial write-up, (longitudinal studies may be an exception if relationships between waves of interviews need special attention for harmonised editing);
- » Use pseudonyms or replacements that are consistent throughout the research team and the project. For example, using the same pseudonyms in publications and follow-up research;
- » Use 'search and replace' techniques carefully so that unintended changes are not made, and misspelt words are not missed;
- » Identify replacements in text clearly, for example with [brackets] or using XML tags such as `<seg>word to be anonymised</seg>`;
- » Create an anonymisation log (also known as a de-anonymisation key) of all replacements, aggregations or removals made and store such a log securely and separately from the anonymised data files.

Example: anonymisation methods in Finland

An example of anonymisation methods is available in the [online version of this guide](#).

Expert tips



1. Data access controls

In situations where (sensitive) personal data are not fully anonymised, data can still be archived and shared by regulating or limiting access to the data. Access controls can permit control down to an individual file level, meaning that mixed levels of access control can be applied to a data collection. You will learn more about choosing the appropriate data access category for your data files in the chapter on archiving and publishing data (see '[Access categories](#)').

2. Irreversible anonymisation

In some countries, anonymisation needs to be irreversible and the original data deleted. Be sure to check the national requirements.

3. Anonymisation tools

The UK Data Archive (n.d.b.) has developed a [Text anonymisation helper tool](#) (downloads in a .zip file) with how to [install instructions](#) via Wiki. It is an add-on MS Word macro for aiding anonymisation of qualitative data.

4. Reading tip

In this [factsheet by OpenAIRE](#) (2017) you are guided in how to balance open access and data protection and advised on what to do when anonymisation isn't possible.

Case study

In a [research study](#) on investigating how couples manage their households during recessions (Gush and Laury, 2015), finding the right balance between confidentiality and usefulness of the data was a [real challenge](#) (UK Data Service, 2017c). Archiving challenges with this project were to anonymise the data and apply optimal access conditions.

Careful judgement was required to apply the level of anonymisation most appropriate for this particular data. The research team members went through the transcripts and removed certain types of identifying data such as names, places of work, and geographic areas. Regarding access conditions, it was decided to make the data available using a [Special Licence](#) (UK Data Service, 2017d; see '[Licensing your data](#)' for other possible licences). Under this kind of licence, a potential user is required not only to register with the UK Data Service, but also to complete a detailed application form and agree to additional restrictions on data handling and usage. The use of the Special Licence then made it possible to apply a minimal level of anonymisation, thus reducing the loss of data quality.

A practice in anonymising qualitative data

Follow the steps to see whether you recognise direct and indirect identifiers in an interview transcript and whether you know how to deal with them accordingly.

Step 1. Read the study background.

Mr Tom Jeavons, aged 63, was suffering from metastatic cancer resulting from a primary site in the bladder. His wife, Sue (58), had been his main carer for many months as he struggled with severe pain, anxiety and other symptoms. Eventually, she received support from the hospice at home team, based at their nearby hospice – St Barbara. 11 days before his death, he was admitted to their inpatient unit, where he died. The case was identified by the staff there as a “critical case”, involving palliative sedation and the difficulties staff experienced in controlling his complex symptoms. Other interviews carried out were with the hospice consultant, Dr Jane O’Connor and three nurses: Elaine McDonald, Claire Smith, and Mark Ferguson. Mr and Mrs Jeavons’ GP, Dr Paul Hyde, was also interviewed which added a different medical perspective, making this an unusual case.

Central themes in all of the interviews were his intractable and distressing symptoms and the repeated requests from Mr Jeavons for euthanasia. His wife mentions earlier discussions with Mr Jeavons about the possibility of going to a Dignitas clinic, but he was already too ill to travel. She also expresses how concerned she was about what Mr Jeavons’s adult children might witness when he was dying in the hospice.

Source: Data collection by Seymour (2010-2012).

Step 2. Read the transcript and uncover direct and indirect identifiers

Read through the interview script and consider what anonymisation would be needed before archiving this transcript for future sharing.

TRANSCRIPT SYMBOLS

INT	Interviewer
RESP	Respondent
[?]	Unintelligible

INT: So, really, it’s as I said to you: I want you to tell me what you can remember about Mr Jeavons’ care in the last week of his life ... or about Mr Jeavons in the last week of his life.

RESP: Yeah, erm, 11 days, Tom was in St Barbara’s Hospice for the last 11 days of his life so...

INT: So if you’d like to talk about that period...

RESP: Yeah.

INT: ...that’d be great.

RESP: Prior to him going in, and we was coping with his care at home, but then he was becoming less and less mobile: he couldn’t go to the toilet; he had a frame, and everything that you added in that was, it was a step to help him but a downward step to the end of how he could cope. We had a Bariatric bed brought into the other room but he insisted in sleeping in his chair. We had St Barbara’s here and, erm, the GP, and, er, we also had him assessed at home as to whether or not we could care for him completely at home. And Tom was about 20-something stone, so he wasn’t easy to manoeuvre and, and the one thing that concerned me was the fact that, erm, they needed four people to move him, you know, if he wanted to go to the toilet or if he wanted to go on a bedpan or anything, and we had the bed in there – which he wouldn’t sleep in. And, erm, basically the, logistically trying to be able to do everything for him and keep him comfortable, we’d have to wait for an on-call four nurses – could be in the middle of the night – and, and sort of the idea of being able to cope, erm, for his safety and wellbeing was, was really compromised. He didn’t want to go into St Barbara’s, he didn’t want to die in hospital, erm, but I just felt I had to take that decision to say, erm, when the guy came out to assess him, erm, he said, ‘We can do it but, you know, you’ve got to say what you’re going to do at three o’clock on Saturday, early hours of

Saturday morning, and he wants to go on the bedpan or you need to change him or whatever.' And, and it, I had to let logic and let my heart... be ruled by my head.

INT: Mm.

RESP: So we got him into St Barbara's., and he went in on the Friday, 11 days before he died, and, erm... when, when he went in – because he couldn't move – from, from a few days before that he wasn't able to move to get to the toilet or anything and we got commodes and things like that and, you know, with having young, young girls in here, we couldn't find him somewhere that he could be private...

INT: Mm.

RESP: ...and that was a bit of a problem for him, because he was a very private man in that, in that way. Erm, so we went into St Barbara's on the Friday and they decided that what they were going to do was going to fit him with a catheter. Well, unfortunately, it was so traumatic for him because all Tom's waterworks had retracted...

INT: Ah.

RESP: ...so much, but there was a determination on the, on the part of the staff to try and make it easier for him to have this catheter put in. Well, it wasn't, it was counter-productive really because, erm, his son came to see his dad, and I was there, and we went out the room and this nurse had spent about an hour and a half trying to get this catheter in. They tried to do it at home, erm, and failed...

INT: Mm.

RESP: ...and of course he was incredibly sensitive, incredibly tender and everything else, and everything had shrunk and retracted so far back it was nigh impossible to actually, to do it without causing him any distress.

INT: Mm.

RESP: So they left it at home but we tried to get it done, erm, in the hospital, they tried to do it, and this lady, erm, had succeeded in getting a catheter in, but he was traumatised by it – there was no other word, he was traumatised – and when myself and his son went back into the room after about an hour and a half, waiting for this thing to, to be finished, er, he actually said to me and to his son, 'Just go away and leave me alone.' And that, unfortunately, was the last time his son saw him, so, Darren lives way over in Seatown. So unfortunate his son's last memory was that. So he stuck with the catheter but the catheter didn't really feel that comfortable, and every time he passed water he was actually yelling in pain. Er, two or three days later they actually took the catheter out and just put him on a pad and, and let him just wee, because, to be honest, did it matter? You know, and to put him through it, he was traumatised with his catheter fitting, and, you know, obviously they're trying to make life easier and more comfortable, erm, but it was, as I say, it was counter-productive.

Anyway, erm... I came home, had a shower, went back in and he was a little bit calmer. Erm... before he went in, erm, he wasn't eating very much or drinking very much, because his, his requirement for food – he kept asking for, for help to die, because he'd enough – he was, he was really, there was no quality; he was in such a lot of pain; he was on such a lot of drugs, and he, he just really, there was no value to him just languishing as he was. Erm, and so it was basically decided that if, if he wanted a drink... a drink would always be there if he wanted one, but there'd be no encouragement, erm, because as, as St Barbara's said, 'We can't kill him,' you know, quite [?], 'We can't...' you know, 'There's nothing we can't... we can keep him out of pain; we can keep him calm, erm, but we can't kill him.' Erm, and I remember him saying to Dr O'Connor 'Just put the boot in, Dr O' Connor.' ... 'Just put the boot...' [?], he'd had enough. Anyway ... [] I cannot criticise the care that they gave him at St Barbara's because it was, you know, fantastic.

Step 3. Have a look at the answers to this exercise

Here you find the answer to what direct and indirect identifiers need to be anonymised. They are underlined and given a number in brackets. At the bottom of the page, you see how anonymisation can be done for each case.

TRANSCRIPT SYMBOLS

INT	Interviewer
RESP	Respondent
[?]	Unintelligible
[]	Edited to maintain anonymity [1 - Added to clarify anonymisation of transcript]

Mr Tom Jeavons **[2 - Delete and replace with [This gentleman]]**, aged 63, **[3 - Delete]** was suffering from metastatic cancer resulting from a primary site in the bladder **[4 - Delete]**. His wife, Sue **[5 - Delete]** (58), **[6 - Delete]** had been his main carer for many months as he struggled with severe pain, anxiety and other symptoms. Eventually, she received support from the hospice at home team, based at their nearby hospice – St Barbara. **[7 - Delete]** 11 days before his death, he was admitted to their inpatient unit, where he died. The case was identified by the staff there as a “critical case”, involving palliative sedation and the difficulties staff experienced in controlling his complex symptoms. Other interviews carried out were with the hospice consultant, Dr Jane O’Connor **[8 - Delete]** and three nurses: Elaine McDonald, Claire Smith and Mark Ferguson **[9 - Delete]**. Mr and Mrs Jeavons’ **[10 - Delete and replace with [The couple’s]]** GP, Dr Paul Hyde, **[11 - Delete]** was also interviewed which added a different medical perspective, making this an unusual case.

Central themes in all of the interviews were his intractable and distressing symptoms and the repeated requests from Mr Jeavons **[12 - Delete and replace with [the patient]]** for euthanasia. His wife mentions earlier discussions with Mr Jeavons **[13 - Delete and replace with [her husband]]** about the possibility of going to a Dignitas clinic, but he was already too ill to travel. She also expresses how concerned she was about what Mr Jeavons’s **[14 - Delete and replace with [his]]** adult children might witness when he was dying in the hospice.

INT: So, really, it’s as I said to you: I want you to tell me what you can remember about Mr Jeavons’ **[15 - Delete and replace with [your husband’s]]** care in the last week of his life ... or about Mr Jeavons **[16 - Delete and replace with [your husband]]** in the last week of his life.

RESP: Yeah, erm, 11 days, Tom **[17 - Delete and replace with [he]]** was in St Barbara’s Hospice **[18 - Delete and replace with [the hospice]]** for the last 11 days of his life so...

INT: So if you’d like to talk about that period...

RESP: Yeah.

INT: ...that’d be great.

RESP: Prior to him going in, and we was coping with his care at home, but then he was becoming less and less mobile: he couldn’t go to the toilet; he had a frame, and everything that you added in that was, it was a step to help him but a downward step to the end of how he could cope. We had a Bariatric bed brought into the other room but he insisted in sleeping in his chair. We had St Barbara’s **[19 - Delete and add [hospice at home]]** here and, erm, the GP, and, er, we also had him assessed at home as to whether or not we could care for him completely at home. And Tom **[20 - Delete and replace with [he]]** was about 20-something stone, so he wasn’t easy to manoeuvre and, and the one thing that concerned me was the fact that, erm, they needed four people to move him, you know, if he wanted to go to the toilet or if he wanted to go on a bedpan or anything, and we had the bed in there – which he wouldn’t sleep in. And, erm, basically the, logistically trying to be able to do everything for him and keep him comfortable, we’d have to wait for an on-call four nurses – could be in the middle of the night – and, and sort of the idea of being able to cope, erm, for his safety and wellbeing was, was really compromised. He didn’t want to go into St Barbara’s **[21 - Delete and replace with [the hospice]]**, he didn’t want to die in hospital, erm, but I just felt I had to take that decision to say, erm, when the guy came out to assess him, erm, he said, ‘We can do it but, you know, you’ve got to say what you’re going to do at three o’clock on Saturday, early hours of Saturday morning, and he wants to go on the bedpan or you need to change him or whatever.’ And, and it, I had to let logic and let my heart... be ruled by my head.

INT: Mm.

RESP: So we got him into St Barbara’s **[22 Delete and replace with [the hospice]]**, and he went in on the Friday, 11 days before he died, and, erm... when, when he went in – because he couldn’t move – from, from a few days before that he wasn’t able to move to get to the toilet or anything and we got commodes and things like that and,

you know, with having young, young girls in here, we couldn't find him somewhere that he could be private...

INT: Mm.

RESP: ...and that was a bit of a problem for him, because he was a very private man in that, in that way. Erm, so we went into St Barbara's **[23 Delete and replace with [the hospice]]** on the Friday and they decided that what they were going to do was going to fit him with a catheter. Well, unfortunately, it was so traumatic for him because all Tom's **[24 - Delete and replace with [his]]** waterworks had retracted...

INT: Ah.

RESP: ...so much, but there was a determination on the, on the part of the staff to try and make it easier for him to have this catheter put in. Well, it wasn't, it was counter-productive really because, erm, his son came to see his dad, and I was there, and we went out the room and this nurse had spent about an hour and a half trying to get this catheter in. They tried to do it at home, erm, and failed...

INT: Mm.

RESP: ...and of course he was incredibly sensitive, incredibly tender and everything else, and everything had shrunk and retracted so far back it was nigh impossible to actually, to do it without causing him any distress.

INT: Mm.

RESP: So they left it at home but we tried to get it done, erm, in the hospital, they tried to do it, and this lady, erm, had succeeded in getting a catheter in, but he was traumatised by it – there was no other word, he was traumatised – and when myself and his son went back into the room after about an hour and a half, waiting for this thing to, to be finished, er, he actually said to me and to his son, 'Just go away and leave me alone.' And that, unfortunately, was the last time his son saw him, so, Darren **[25 - Delete and replace with [his son]]** lives way over in Seatown **[26 - Delete and replace with [he lives some distance away]]**. So unfortunate his son's last memory was that. So he stuck with the catheter but the catheter didn't really feel that comfortable, and every time he passed water he was actually yelling in pain. Er, two or three days later they actually took the catheter out and just put him on a pad and, and let him just wee, because, to be honest, did it matter? You know, and to put him through it, he was traumatised with his catheter fitting, and, you know, obviously they're trying to make life easier and more comfortable, erm, but it was, as I say, it was counter-productive.

Anyway, erm... I came home, had a shower, went back in and he was a little bit calmer. Erm... before he went in, erm, he wasn't eating very much or drinking very much, because his, his requirement for food – he kept asking for, for help to die, because he'd enough – he was, he was really, there was no quality; he was in such a lot of pain; he was on such a lot of drugs, and he, he just really, there was no value to him just languishing as he was. Erm, and so it was basically decided that if, if he wanted a drink... a drink would always be there if he wanted one, but there'd be no encouragement, erm, because as, as St Barbara's **[27 - Delete and replace with [the hospice]]** said, 'We can't kill him,' you know, quite [?], 'We can't...' you know, 'There's nothing we can't... we can keep him out of pain; we can keep him calm, erm, but we can't kill him.' Erm, and I remember him saying to Dr O'Connor **[28 - Delete and replace with [the doctor]]** 'Just put the boot in, Dr O' Connor **[29 - Delete and replace with [doctor]]** ! ... 'Just put the boot...' [?], he'd had enough. Anyway ... [] I cannot criticise the care that they gave him at St Barbara's **[30 - Delete and replace with [the hospice]]** because it was, you know, fantastic.

5.6 Copyright

Copyright is an internationally recognised form of intellectual property right, which arises automatically as a result of original work such as research. It does not need to be registered to apply to a piece of work.

Copyrighted output from research could include spreadsheets (and other forms of originally selected and organised data), publications, reports and computer programs. Copyright will not cover the underlying facts, ideas or concepts, but only the particular way in which they have been expressed. The right will lie with the author of the work, or with their relevant institution—different universities will have different policies on intellectual property.

A copyrighted work cannot usually be published, reproduced, adapted or translated without the owner's permission.

Key copyright considerations for researchers

Whether you want to reuse someone else's data or if you are planning to archive and share your own, you should ask yourself who the copyright holder of the datasets is (also see '[Licensing your data](#)'). Are you allowed to use them and in what way? Are you allowed to archive and publish them in a data repository? How do you answer the question who the copyright holder of a dataset is? Is it you, your employer, the data archive, fellow researchers? The answer depends on multiple factors, such as who had input into creating the research data, whether data were used from other datasets, and what the researcher's contract of employment stipulates.

Key copyright considerations for researchers are highlighted below:

Joint ownership

In two cases multiple copyright holders exist and joint ownership is implied:

» **Datasets created by multiple researchers**

When data is collected, and created by multiple researchers, then multiple researchers may be listed as joint copyright holders, with all gaining and retaining intellectual property rights. Prior to archiving it is important to ensure permission for depositing data is given by all copyright holders as well as participants.

» **Derived datasets**

The key issue with derived data is the matter of copyright ownership. Because the resulting data is derived from previously created data the permission of the original copyright holder should be sought before the data is deposited with a repository. The best practice is for researchers to try to negotiate the sharing of derived data with the data suppliers at the time of acquisition or purchase. If permission is granted then they should also be listed as a joint copyright owner.

Database rights

Database rights acknowledge the investment made by a researcher in developing a database, even in cases where this does not involve a creative aspect. The organisation, structuring of a database and the selecting of which data to include in the database are all decisions which can receive protection through copyright legislation. If you want to use (large parts of) a database you should always ask the permission of the database creator.

Provisions in a contract

Depending on the employer you might have a stipulation in your contract that any works which are created during employment are the intellectual property of the employer. But, even where contracts of employment make this statement we typically find that the employer is happy for the researcher to be listed as the copyright holder or for the employer and employee to be listed as a joint holder.

Repository copyright rules

Most repositories operate a system of not acquiring any copyright ownership in the data*. Before you deposit your data the repository will need to be informed – and confirm – who the copyright owner is.

What we typically see in practice is that the researcher who authors (creates) the work is listed as the copyright owner for the dataset when it is deposited in a repository. Repositories act merely as a facilitator of access to the data, with some guaranteeing to curate and provide permanent access to the data.

* However, this can differ from country-to-country and archive-to-archive. Researchers should, therefore, clarify with the repository copyright rules before depositing the data, and ideally before conducting research, so that consent forms and information sheets can inform participants accurately of who will own – and have copyright of – the data once the research project has commenced.

Case studies

In the following case studies, you can identify the potential copyright issues and state how you would address these in practice.

Case Study 1 – Copyright of Archived Data

A researcher uses [International Social Survey Programme \(ISSP, n.d.\)](#) data obtained from ZACAT/GESIS - Leibniz Institute for the Social Sciences in Germany. These data are freely available to registered users. The researcher incorporates some of the ISSP data within a database containing his own research data. Can this database be deposited with another archive?

Answer: Although the ISSP data are available for free to all researchers, this does not mean that the data can be published in another archive and made available to others. The data can be incorporated into a database and used for personal analysis. But, before this dataset can be deposited with another archive, permission must be sought from the owner of the original data.

Case Study 2 – Copyright of Data in the Public Domain

A researcher studies how health issues around obesity are reported in the media in the last 10 years. Freely available newspaper websites and library sources are used to obtain articles on this topic. Articles or excerpts are copied into a database and coded according to various criteria for content analysis. (i) Can the researcher use such public data without breaching copyright? (ii) Can the database be archived and shared with other researchers?

Answer: Even though the articles are publicly available, they are still under copyright. Whilst such information can be used for personal research purposes (e.g. in the UK this would be under the broad exemption of 'fair dealing'), the articles cannot be archived unless permission is obtained from the newspapers; otherwise this would breach copyright.

Case Study 3 – Copyright of Survey Questions

A researcher wishes to reuse a set of questions from an existing survey questionnaire, to compare results between the newly proposed survey and the original.

Answer: The survey questions and instruments will be copyright protected, with copyright residing with the organisation who commissioned, designed or conducted the survey (unless the original creator/owner transfers all ownership rights). The researcher needs to contact the copyright holder directly for permission to reproduce the questionnaire text for any new use. Some questionnaires will contain measurement scales, batteries of questions or classifications. These instruments are again copyrighted. Therefore, to reproduce them the researcher will need permission.

Case Study 4 – Copyright of Interviews with Stay-at-Home Parents

A researcher interviews various stay-at-home parents about their careers and produces audio recordings and near-verbatim transcripts herself. The researcher analyses this material and offers it to a data archive. The researcher did not get signed copyright transfers for the interviewees' words. What are the rights issues surrounding this offer of data?

Answer: In this case, the stay-at-home parents hold copyright in their own recorded words, whilst the researcher holds copyright over the transcribed interviews. Quoting large extracts of the data, either in publications or by archiving the transcripts, would breach the copyright of the interviewees in their recorded words. If the researcher wants to publish large extracts of data, or archive the transcripts, they need to request permission to do so from the interviewees or request that the interviewee transfers the copyright of the interview content to the researcher, which could be achieved through the use of a Recording Agreement.

5.6.1 Diversity in copyright

Copyright legislation is created at an individual country level as there is not an international copyright law, though many countries have signed up to the [Berne Convention](#) (WIPO, 1979). For these reasons, it is important that researchers identify the relevant national copyright legislation.

The European Commission is looking to reform EU copyright law further, having published a package of reform proposals, which currently include a Directive and a Regulation. One of the aims of these reforms is to improve copyright rules on research and education.

Below we list information on national copyright legislation, what is covered, the copyright duration and the exceptions and limitations.

For updated information, also see the [online version of this guide](#).

Croatia

National legislation

Copyright and Related Rights Act and the Act on Amendments to the Copyright and Related Rights Act (Official Gazette N°167/2003, N°79/2007, N°80/2011, N°141/2013, N°127/2014, N°62/2017, N°96/2018)

[\[Croatian\]](#) [\[English\]](#)

Duration

Article 99: "Copyright shall run for the life of the author and for 70 years after his death, irrespective of the date when the work is lawfully released, unless otherwise provided by the Copyright Act."

Article 152: "Rights of a producer of a database shall run for 15 years as from the date of the completion of the making of the database. If the database is lawfully disclosed during this period, the rights shall run for 15 years as from the first such disclosure."

What is covered

The Copyright and Related Rights Act regulates:

copyright - rights of authors in respect of their works in the literary, scientific and artistic domains; and related rights, among which the rights of producers of databases in respect of their databases might be of relevance to researchers.

Article 5: "A copyright work shall be an original intellectual creation in the literary, scientific and artistic domain, having an individual character, irrespective of the manner and form of its expression, its type, value or purpose."

Article 7: "Collections of independent works, data or other materials, such as encyclopaedias, collections of documents, anthologies, databases, and the like, which by reason of the selection or arrangement of their constituent elements constitute personal intellectual creations of their authors shall be protected as such. Databases, under this Act, shall be collections arranged according to certain system or method, the elements of which are individually accessible by electronic or other means."

Chapter 6: "Rights of producers of databases, Article 147: A database, under this Chapter of the Act, shall mean a collection of independent works, data or other materials in any form, arranged in a certain systematic or methodical way and individually accessible by electronic or other means, whereby either the obtaining, verification or presentation of its contents requires a qualitatively and/or quantitatively substantial investment in terms of resources, time and efforts engaged."

Exceptions and limitations

Unprotected creations

Article 8: "(1) The subject matter of copyright shall include expressions and not ideas, procedures, methods of operation or mathematical concepts as such. (2) The subject matter of copyright shall not include: 1. discoveries, official texts in the domain of legislation, administration, judiciary (acts, regulations, decisions, reports, minutes, judgments, standards, and the like) and other official works and their collections, disclosed for the purpose of officially informing the public; 2. news of the day and other news, having the character of mere items of press information; (3) Folk literary and artistic creations in their original form shall not be the subject matter of copyright, but their communication to the public is subject to the payment of remuneration, as for the communication to the public of protected copyright works. The remuneration shall be the revenue of the budget, and shall be used for improving the creativity in the field concerned."

Exceptions regarding rights of producers of databases

Article 150: "An authorized user of a disclosed database may, without the authorization of its producer, use the substantial parts of its contents in the case: 1. referred to in Article 149, item 1 of this Act for private use of a non-electronic database; 2. referred to in Article 149, item 1 of this Act for use intended for teaching or scientific research, provided that the source is indicated and to the extent justified by the non-commercial purpose; 3. referred to in Article 149, items 1, 2, 3, and 4 of this Act for use required for public safety, or for administrative or judicial proceedings."

Czech Republic

National legislation

[Copyright Act No. 121/2000; 07/2017](#) (2000) (in Czech).

In English, version 01/2015: Consolidated text of Act No. 121/2000 on Copyright and Rights Related to Copyright and on Amendment to Certain Acts ([the Copyright Act](#) (Ministry of Culture Czech Republic (2000); pdf, downloads on click)).

Duration

Copyright duration varies based on the type of work. Unless stipulated otherwise, economic rights shall run for the life of the author and 70 years after his death but this is 50 years for Performer's Economic Rights, phonogram, broadcasters, and publishers. The right sui generis of the maker of the database shall run for 15 years from the making of the database.

What is covered

- » Vol.I., Art.2: (1)"...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.. A work shall be, without limitation, a literary work expressed by speech or in writing, a musical work, a dramatic work or musical-dramatical work, a choreographic work and pantomimic work , a photographic work and a work produced by a process similar to photography, an audiovisual work such as a cinematographic work, a work of fine arts such as a painting, graphic or sculptural work, an work of architecture including an urban design work, a work of applied art, and a cartographic work.
- » A computer program shall also be considered a work if it is original in the sense that it is the author's own intellectual creation. 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. No other criteria shall be applied to determine their eligibility for that protection. A photograph or a work produced by a process similar to photography, which are original in the sense of the first sentence, shall be protected as a photographic work.

- » A work which is the outcome of the creative adaptation of another work, including its translation into another language, shall also be subject to copyright. This shall be without prejudice to the rights of the author of the adapted or translated work.
- » A collection like a journal, encyclopaedia, anthology, exhibition, or any other collection of independent works or other elements that by reason of their selection and of the arrangement of the content meet the conditions set out in Paragraph 1 above, is a collection of works.
- » The items that are not works hereunder, shall include, but are not limited to the theme (subject) of a work as such, the news of the day and any other fact as such, an idea, procedure, principle, method, discovery, scientific theory, mathematical and similar formula, statistical diagram and similar item as such”.

Exceptions and limitations

No copyright, Art. 3: a) an official work, such as a legal regulation, decision, public charter, publicly accessible register and collection of its documents, and also any official draft of an official work and other preparatory official documentation including the official translation of such work, Chamber of Deputies and Senate publications, a memorial chronicle of a municipality (municipal chronicles), a state symbol and symbol of a municipality, and any other such works where there is public interest in their exclusion from copyright protection, b) creations of traditional folk culture, unless the real name of the author is commonly known and the works are anonymous or pseudonymous (Article 7); such works may only be used in a way that shall not detract from their value.

Finland

National legislation

[Copyright Act](#) (404/1961, amendments up to 608/2015) (2015).

Duration

Copyright shall subsist until 70 years have elapsed from the year of the author's death or from the year of death of the last surviving author.

What is covered

A person who has created a literary or artistic work shall have copyright therein, whether it be a fictional or descriptive representation in writing or speech, a musical or dramatic work, a cinematographic work, a photographic work or other work of fine art, a product of architecture, artistic handicraft, industrial art, or expressed in some other manner. Maps and other descriptive drawings or graphically or three-dimensionally executed works and computer programs shall also be considered literary works.

Exceptions and limitations

There shall be no copyright: 1) in laws and decrees; 2) in resolutions, stipulations and other documents which are published under the Act on the Statutes of Finland (188/2000) and the Act on the Regulations of Ministries and other Government Authorities (189/2000); 3) treaties, conventions and other corresponding documents containing international obligations; 4) decisions and statements issued by public authorities or other public bodies; 5) translations of documents referred to in paragraphs 1–4 made by or commissioned by public authorities or other public bodies.

The provisions of subsection 1 shall not apply to independent works contained in the documents referred to in the subsection.

Germany

National legislation

[Act on copyright and related rights](#) (2016).

Duration

Copyright expires 70 years after the author's death.

What is covered

Protected Works

Protected works in the literary, scientific and artistic domain include, in particular:

1. Literary works, such as written works, speeches, and computer programs;
2. Musical works;
3. Pantomimic works, including works of dance;
4. Artistic works, including works of architecture and of applied art and drafts of such works;
5. Photographic works, including works produced by processes similar to photography;
6. Cinematographic works, including works produced by processes similar to cinematography;
7. Illustrations of a scientific or technical nature, such as drawings, plans, maps, sketches, tables and three-dimensional representations.

Only the author's own intellectual creations constitute works within the meaning of this Act.

Collections and database works

1. Collections of works, data or other independent elements which by reason of the selection or arrangement of the elements constitute the author's own intellectual creation (collections) are protected as independent works without prejudice to an existing copyright or related right in one of the individual elements;
2. A database work within the meaning of this Act is a collection whose elements are arranged systematically or methodically and the individual elements are individually accessible by electronic or other means. A computer program (section 69a) used in the creation of the database work or to provide access to its elements does not constitute an integral part of the database work.

Exceptions and limitations

Official works

1. Acts, statutory instruments, official decrees and official notices, as well as decisions and official head notes of decisions, do not enjoy copyright protection;
2. The same applies to other official texts published in the official interest for general information purposes, subject to the proviso that the provisions concerning the prohibition of alteration and the indication of sources in section 62 (1) to (3) and section 63 (1) and (2) shall apply mutatis mutandis.

Authors in employment or service

The provisions of this subchapter shall also apply where the author has created the work in the fulfilment of obligations resulting from an employment or service relationship unless otherwise provided in accordance with the terms or nature of the employment or service relationship.

The German 'Gesetz über Urheberrecht und verwandte Schutzrechte (UrhG)' makes several exceptions when works are used in the context of research or teaching.

Greece

[Law 2121/1993](#), as in force for copyright issues;

[Law 4481/2017](#) for copyright collective management issues.

Duration

Copyright expires 70 years after the author's death.

What is covered

Protected Works

Protected works in the literary, scientific and artistic domain include (described in the text of the law as any original intellectual literary, artistic or scientific creation, expressed in any form), notably:

- » written or oral texts,
- » musical compositions with or without words,
- » theatrical works accompanied or unaccompanied by music,
- » choreographies and pantomimes,
- » audiovisual works,
- » works of fine art, including drawings, works of painting and sculpture, engravings and lithographs,
- » works of architecture and photographs,
- » works of applied art,
- » illustrations, maps and three-dimensional works relative to geography, topography, architecture or science.

Only the author's own intellectual creations constitute works within the meaning of this Act.

Collections and database works

1. The term work also covers translations, adaptations, arrangements and other alterations of works or of expressions of folklore, as well as collections of works or collections of expressions of folklore or of simple facts and data, such as encyclopaedias and anthologies, provided the selection or the arrangement of their contents is original;
2. Databases which, by reason of the selection or arrangement of their contents, constitute the author's intellectual creation, are as such by copyright. The copyright protection shall not extend to the contents of databases and shall be without prejudice any rights subsisting in those contents themselves. Database is a collection of independent works, data or other, materials arranged in a systematic or methodical way and individually accessible by electronic or other means.
3. Computer programs and their preparatory design material are literary works within the meaning of the provisions on copyright protection. Protection in accordance with this Law applies to the expression in any form of a computer program. Ideas and principles which underlie any element of a computer program, including those which underlie its interfaces, are not protected under this Law. A computer program is protected if it is original in the sense that it is the author's personal intellectual creation.

Exception and Limitations

Copyright protection does not apply to:

- » official texts expressive of the authority of the State, notably to legislative, administrative or judicial texts
- » expressions of folklore,
- » news information
- » simple facts and data.

The Greek copyright law provides for several exceptions when works are used in the context of research or teaching in articles 18-28 C of law 2121/1993 - see at <https://www.opi.gr/en/library/law-2121-1993#ch4>.

Netherlands

National legislation

[Auteurswet](#) (2017) (in Dutch).

Duration

Copyright is in force until 70 years after the death of the author.

What is covered

The Copyright Act provides protection for works that display a certain creativity or originality.

Exceptions and limitations

Bare facts as such, citations and works produced by the official authorities do not fall under The Copyright act.

North Macedonia

National legislation

Law on copyrights and related rights (Official gazette of the R. Macedonia No. 115/10, 140/10 и 51/11).

What is covered

This law regulates the copyrights of authors over their work, among others, the rights of "...authors of data sets over their works, or related rights...", as well as the practicing and protection of copyrights and related rights (Article 1). According to this law, related rights can also be "databases and their authors". Related rights are regulated with the General provisions on related rights, especially in Part 6 - The rights of authors of databases (Article 118-122). The law includes the following aspects: definition of database authors; contents of the rights of the authors of databases; the scope of protection; rights and obligations of the legal users; restriction of the rights; and duration of the rights of authors of databases. The law also foresees the possibility of regulation of collective management of related rights.

Norway

The new Norway Copyright Act of 16.06.2018 amends the previous one drafted in 1961 (Act No. 2 of May 12, 1961) and consolidated in 2015. Information about the main normative aspects is available [here](#).

Duration

Copyright is in force until 70 years after the death of the author.

Database rights are in force for 15 years after the year of production or after the year of the last update.

What is covered

Research data are rarely protected by copyright, but more often it may be protected by database rights.

A simple rule of thumb is that Copyright protects original works of authorship (scientific articles, books, reports, blogs) but not facts, raw data, etc. unless they are selected and arranged in an original way (cf. Lov om opphavsrett til åndsverk/2018-06-16/§2). One might think, for example, that the selection of variables or other data is unique or creative, but if your selection is motivated by rational questions or objective considerations, it is not "creative" in a copyright sense.

The Norwegian Copyright Act covers Copyright and Related Rights (Neighboring Rights), Enforcement of IP and Related Laws, IP Regulatory Body and Industrial Designs.

The Copyright owner is the author or person to whom rights have been transferred (e.g. the publisher); cf. Lov om opphavsrett til åndsverk/2018-06-16/§8.

According to §41 in Norwegian law, Database right protects databases that are the result of a substantial investment in either the obtaining, verification or presentation of its contents. This means that the investment in the creation of the contents does not give you database rights.

The database right belongs to the maker of the database, i.e. the person (natural or legal) who bears the risk of the investment. Thus, if a database is created in the course of employment, the right will belong to the employer, and not the employee(s). See Lov om opphavsrett til åndsverk/2018-06-16/§41.

Exceptions and limitations

§ 14 of the Norwegian Copyright Act states that laws, regulations, court decisions and other decisions by public authorities are not Copyright protected. The same applies to proposals, investigations and other statements relating to public authority exercise and has been issued by the public authority, publicly appointed council or selection or published by the public.

§ 1-4 of the Norwegian Copyright Act is an exception which allows the production of a copy for research purposes. This provision is relatively new and allows research institutions to apply to the Norwegian Ministry of Culture for the permission to produce copies for research purposes, for example, to cover special needs in language research. The condition is that the production of copies shall not lead to a proliferation in violation of the rights holder's interests and the copy of the product must otherwise conflict with the rights owner's own exploitation of the work.

Serbia

National legislation

The Law on Copyright and Related Rights (Official gazette of RS 104/2009, 99/2011 from 27.12.2011 and 119/2012)

Duration

Pecuniary rights shall last for the life of an author and 70 years after his/her death. The moral rights of an author shall last even after the expiration of his/her pecuniary rights. Co-authors' pecuniary rights shall expire after 70 years elapse from the death of the author that was the last to die. Pecuniary rights concerning the work whose author is unknown (anonymous work or work under a pseudonym) shall expire after 70 years elapse from the date of its disclosure. Should its author reveal his/her identity before the expiration of such term, the pecuniary right shall last the same as if its author's identity has been known since the date of its disclosure. Copyright on the collective works lasts for 70 years from the date of the legal publication of the work.

What is covered

A work of authorship, in particular: Written works (e.g. books, brochures, articles, translations, computer programs in any form of their expression, including their preparatory design material and other); Spoken works (lectures, speeches, orations, etc.); Dramatic, dramatic-musical, choreographic and pantomime works, as well as works originating from folklore; Works of music, with or without words; Films (cinema and television); Fine artworks (paintings, drawings, sketches, graphics, sculptures, etc.); Works of architecture, applied art and industrial design; Cartographic works (geographic and topographic maps); Drawings, sketches, dummies and photographs; and The direction of a theatre play (Article 2). A collection of the works of authorship, which in view of the selection and arrangement of its integral parts, meets the requirements referred to in Article 2, (an encyclopedia, collection of works, anthology, selected works, music collection, photograph collection, graphic map, exhibition and the like), shall also be deemed a work of authorship; a collection of folk literary and artistic creations, as well as a collection of documents, court decisions and similar materials, which in view of their selection and arrangement. A collection

shall also be understood to mean a database, regardless of whether it is in a mechanically or otherwise legible form, which in view of the selection and arrangement of its integral parts.

Exceptions and limitations

The protection of copyright shall not apply to general ideas, procedures and methods of operations or mathematical concepts as such, as well as concepts, principles and instructions included in a work of authorship; Laws, decrees and other regulations; Official materials of state bodies and bodies performing public functions; Official translations of regulations and official materials of state bodies and bodies performing public functions; Submissions and other documents presented in the administrative or court proceedings.

Slovenia

National Legislation

[Copyright and Related Rights Act](#)

The Act regulates: the right of authors with respect to their works of literature, science and art (copyright); the rights of performers, producers of phonograms, film producers, broadcasting organizations, publishers and makers of databases (related rights).

What is covered

As copyright works are considered in particular:

1. spoken works such as speeches, sermons, and lectures;
2. written works such as belletristic works, articles, manuals, studies, and computer programs;
3. musical works with or without words;
4. theatrical or theatrical-musical works, and works of puppetry;
5. choreographic works and works of pantomime;
6. photographic works and works produced by a process similar to photography;
7. audiovisual works;
8. works of fine art such as paintings, graphic works, and sculptures;
9. works of architecture such as sketches, plans, and built structures in the field of architecture, urban planning, and landscape architecture;
10. works of applied art and industrial design;
11. cartographic works;
12. presentations of a scientific, educational or technical nature (technical drawings, plans, sketches, tables, expert opinions, three-dimensional representations, and other works of similar nature).

Sweden

National legislation

The [Act on Copyright in Literary and Artistic Works](#) (1960:729).

Duration

Copyright of a work shall subsist until the end of the seventieth year after the year in which the author deceased. If a work has two or more authors whose contributions do not constitute independent works, the copyright shall belong to the authors jointly. In that case, copyright subsists until the end of the seventieth year after the year in which the last surviving author deceased.

What is covered

- » Anyone who has created a literary or artistic work shall have copyright in that work, regardless of whether it is:
- » A fictional or descriptive representation in writing or speech;
- » A computer program;
- » A musical or dramatic work;
- » A cinematographic work;
- » A photographic work or another work of fine arts;
- » A work of architecture or applied art;
- » A work expressed in some other manner.

Maps and other works of a descriptive nature executed as drawings, engravings, or in a three-dimensional form, shall be considered as literary works. What is prescribed in this Act concerning computer programs shall mutatis mutandis apply also to preparatory design material for computer programs. (Act 1994:190)

Exceptions and limitations

- » Copyright does not subsist in:
- » Laws and other regulations;
- » Decisions by public authorities;
- » Reports by Swedish public authorities;
- » Official translations of texts mentioned under 1–3.

However, copyright subsists in works of the following kinds when they form part of the following documents: maps, works of drawing, painting or engraving, musical works or works of poetry. (Act 2000:92).

Switzerland

National legislation

[Swiss Federal Act on Copyright and Related Rights](#) (1992). The [DICE project](#) (Université de Genève, 2010) provides a useful handbook on copyright in education in Switzerland.

UK

National legislation

The Copyright, Design and Patents Act dates from 1988.

Duration

Copyright duration varies based on the type of work. For literary and artistic works it is 70 years from the end of the year of the death of creator, for sound recordings it is 50 years from date of creation and for typographical arrangements, it is 25 years from date of publication. For Crown Copyright the duration can be 50 years from the date of publication or 125 years from the date of creation.

What is covered

Original literary, dramatic, musical or artistic works, sound recordings, films, broadcasts or cable programmes or the typographical arrangement of publications.

Exceptions and limitations

The UK has created various exceptions including 'fair dealing' and 'non-commercial research and private study'.

Obstacles to the trans-European archiving and sharing of research data

Making research data as openly available as possible is a widely recognised goal. For researchers working on an interdisciplinary project involving several countries, it can be difficult to fully comprehend in which ways open access to research data can be legally obtained. European national laws still diverge.

[A report from Knowledge Exchange](#) (Knowledge Exchange, 2011) concludes that it will remain difficult to predict when particular files of research data are protected because of:

» Diversity in copyright protection

Even though most research data will fail to meet the criteria for copyright protection because they are not likely to be considered as "works" (they mainly concern facts), the lack of harmonisation of the criteria for copyright protection in Europe is tricky. E.g., whereas Germany, Denmark, and the Netherlands have a relatively similar (higher) originality standard, the UK has a very low standard (skill, judgment, and labour) making it possible that collections of research data are easily granted full copyright protection.

» Diversity in copyright owner

If protection applies, the right holder's consent is required for sharing the data. However, the designation of the copyright owner is also different in different jurisdictions. Although in many cases the maker of the work will be considered to be the author and therefore the right holder, only Dutch and UK law designate the employer as the right holder if the work was made in the course of employment.

Licences as a way forward

Therefore, the authors conclude that to ensure that research data can be shared and reused freely licences should always be obtained from the potential rights holders. With the right licence, researchers can waive claims to any IP rights that might apply to research data that they generate in the course of publicly funded research. In the chapter '[Archiving and publishing data](#)' we will look into '[Data licensing](#)'.

5.7 Adapt your DMP: part 5



This is the fifth of seven 'Adapt your DMP' sections in this tour guide.

After working on this chapter, you should be able to define your strategy in protecting the rights of your participants whilst making your data available for as full and effective use as possible for the scientific community and the public. Filling in this part of your DMP will show how you are taking legal and ethical factors into consideration and it can actually help you navigate ethical review (self-assessment

or formal).

To adapt your DMP, consider the following elements and corresponding questions:

Type of data

- » Are you collecting personal data or do your data in any other way require special protection?
- » How is (sensitive) personal data protected during the project? (also see [Storing your data](#))

Ethical review (if applicable)

- » Does your project require approval by a local ethics committee?
- » If so, how are you preparing for this review?
- » How will possible ethical issues be taken into account, and codes of conduct followed?

Informed consent (if applicable)

- » Do you require informed consent for your project? If so, how will permission be obtained?
- » Are you gaining written consent from respondents to share data beyond your research?
- » Did you discuss data archiving and sharing with the respondents from whom you collected the data?
- » Are you adequately documenting your consent to comply with the GDPR?
- » How are consent files organized and stored?

Protecting participants

- » How are you going to protect the privacy of your participants?
- » Are you complying with data protection legislation?
- » Do you need to anonymise data, for example, to remove identifying information or personal data, during research or in preparation for sharing?

Intellectual property

- » Are there IPR or copyright issues to consider?
- » Have you established who owns the copyright in your data? Might there be joint copyright?
- » Will permission be needed to collect/reuse the data?
- » Will these rights be transferred to another organisation for data archiving?

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 6

Archive & Publish

Contents

Main take-aways	156
6.1 Towards archiving & publication	157
6.2 Selecting data for publication	159
6.3 Data publishing routes	160
6.4 Publishing with CESSDA archives	164
6.4.1 Citing your data	165
6.4.2 Licensing your data	167
6.4.3 Access categories	169
6.5 Promoting your data	172
6.6 Adapt your DMP: part 6	174
Sources and further reading	175

[View the online version of this chapter](#)

Main authors of this chapter

Sonja Bezjak, Slovenian Social Science Data Archives (ADP)

Irena Vipavc Brvar, Slovenian Social Science Data Archives (ADP)

Introduction



High-quality data have the potential to be reused in many ways. Archiving and publishing your data properly is at the core of making your data FAIR and will enable both your future self as well as others to get the most out of your data.

In this chapter, we venture into the landscape of research data archiving and publication. We will guide you in making an informed decision on where to archive and publish your data in such a way that others can properly access, understand, use and cite them.

Main take-aways

After completing your journey through this chapter on archiving and publishing data, you should:

- » Understand the difference between data archiving and data publishing;
- » Be aware of the benefits of data publishing;
- » Be able to differentiate between different data publication services (data journal, self-archiving, a data repository);
- » Be able to select a data repository which fits your research data's needs;
- » Be aware of ways to promote your research data publication;
- » Be able to answer the [DMP questions](#) which are listed at the end of this chapter and adapt them to your own DMP.

6.1 Towards archiving & publication

This chapter in the tour guide is all about securing your research data's future for the following purposes:

» **Archiving data for future reference**

Research data archiving is about storing and preserving research data for the long term. When you archive your data, you make sure you can read and access the data later on. You can then also allow access to others for verification purposes when such a request arrives. In all cases, you should store your data safely, in a suitable file format, with adequate documentation.

» **Publishing data for reuse**

To make your data reusable for purposes beyond the one for which you collected them, you should publish your data. Publishing your data is the act of publicly disclosing the research data you have collected, making them findable, accessible and reusable.

The incentives that motivate interviewed researchers to archive and publish their research data (Van den Eynden & Bishop, 2014; Hahnel et al., 2017) fall into four main categories:

Career benefits

Data publication may lead to increased visibility, reuse and citation and therefore recognition of scholarly work.

A number of studies show the impact of data publication on citation rates. Articles for which the underlying data is published are more frequently cited than articles for which this is not the case. Studies from social science (Pienta, Alter & Lyle, 2010), genetics (Piwowar and Vision, 2013; Botstein, 2010), astronomy (Henneken and Accomazzi, 2011; Dorch 2012) and oceanography (Sears 2011, Belter 2014) confirm this effect.

Be aware that whenever you use the published data you are obliged to cite them. For more information see the [paragraph on data citation](#).

Scientific progress

Data archiving and publication has direct benefits for the research itself (more robust), for the discipline and for science in general by enabling new collaborations, new data uses and establishing links to the next generation of researchers.

A [tweet](#) (ESS ERIC, 2017) from the [European Social Survey](#) (n.d.) is just one of the many, many examples of how sharing high-quality datasets leads towards new insights. The European Social Survey is widely accessible and used by many researchers.

Norms

Norms of the project, research group, and/or discipline may determine whether a researcher is prone to publish his/her data. Overall, the openness of research data is at the heart of scientific ethics as is illustrated by the quote below.

Sociologists make their data available after completion of the project or its major publications, except where proprietary agreements with employers, contractors, or clients preclude such accessibility or when it is impossible to share data and protect the confidentiality of the data or the anonymity of research participants (e.g., raw field notes or detailed information from ethnographic interviews) | American Sociological Association (1999).

External drivers

External drivers like research data management policies from research funders and publishers have a significant influence on data archiving and publication:

» Funders

Some funders consider costs related to data archiving and publication eligible and require a DMP. For a list of funder requirements see the [‘European diversity in funder requirements’](#) section of this tour guide.

» Publishers

Scientific journals are increasingly adopting data availability policies that advise or even request authors of manuscripts to make the research data, on which a manuscript is based, available. For example, PLOS One says in its data availability statement:

All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository unless already provided as part of the submitted article. Repositories may be either subject-specific (where these exist) and accept specific types of structured data, or generalist repositories that accept multiple data types, such as Dryad | [PLOS One](#) (2014a).

In the coming paragraphs, the main focus will be on securing high-quality datasets for the future by combining data archiving and data publishing.

6.2 Selecting data for publication

Should you publish your data or shouldn't you? And if so, which part of it? Sometimes this question is rather straightforward to answer. For example, due to research funders' demands. Or because you strongly believe it is not up to you to withhold any latent knowledge hidden in your data from future researchers.

Even so, not all data are created equal and data publishing does involve an investment of time and money. Some datasets have a more obvious reuse potential than others. By following the arguments below (Tjalsma and Rombouts, 2011) you can see for yourself whether (part of) your dataset is an obvious candidate for data publication.

» **Does your dataset have reuse potential?**

Does your data have potential value in terms of reuse, national/international standing and quality, historical importance, uniqueness (the data contain non-repeatable observations), originality, size, scale, costs of data production or innovative nature of the research? Could you foresee that secondary analyses on your data would benefit science? If your answer to any of these is yes, your dataset has serious reuse potential.

» **Is your dataset reusable?**

To be suited for re-use your dataset should be functionally usable. Can the data be read and used? Are metadata available and are they sufficient to enable future users to understand your data? Are there any legal objections which prevent the data from being published? If you are sure you got these practicalities covered and your data have potential value in terms of reuse, you are 'good to go'.

If your data are not 'good to go' at this point but do have re-use potential, do not worry. Most of the time it is not too late to document your data properly and address other issues like gaining consent for sharing retrospectively (see ['Informed Consent'](#)).

6.3 Data publishing routes

It is expected that a Data Publication will ensure that data will potentially be considered as a first-class research output
| Knowledge Exchange (2013).

For a dataset to “count” as a publication, it should follow a similar publication process to an article (Brase et al., 2009) and should be:

- » Properly documented with metadata;
- » Reviewed for quality;
- » Searchable and discoverable in catalogues (or databases);
- » Citable in publications.

The authors of a report from Knowledge Exchange (Knowledge Exchange, 2013) define this type of data publication as ‘Publishing with a capital P’ and compare it with ‘publishing with a small p, meaning that researchers publish their data files on a website somewhere. Publishing with a small “p” means that there are no guarantees that the data will be there after some time or that the files will not get corrupted.

Five routes

There are different ways to publish your data. Your preference may depend on the existing practices in your discipline or on the expectations of your funder.

According to a [survey by Wiley](#) (2014), the preferred way of publishing data is as supplementary material of a journal article. That may change as more data repositories become available, and more scientific journals recommend depositing in them. A data repository is a digital archive collecting, preserving and displaying datasets, related documentation, and metadata ([OpenAIRE, 2017](#))

In the comparisons below we show five ways of publishing your data, together with their advantages and disadvantages.

Journal supplementary material service

Advantages

- » Most likely to comply with the journal or publisher’s requirements;
- » Data readily available alongside published findings.

Disadvantages

- » May be costly;
- » May claim copyright over the data;
- » May keep data behind a subscription wall;
- » Unlikely to offer a data repository’s functionality or long-term solution;
- » May not apply user-friendly or preservation formats;
- » More likely to accept subsets rather than complete datasets.

Institutional data repository

Advantages

- » Most likely to accept any data of value, especially if no suitable home can be found for it elsewhere, and to ensure that policy requirements for long-term access are met;
- » Researchers may trust such a repository more readily;
- » Possibly no charge for the data deposit;
- » May make your data visible via dissemination and promotion.

Disadvantages

- » May not offer sustainable long-term access to your data collection;
- » Might not have sufficient expertise in data and metadata standards needed for long time preservation and access.

General purpose repository

Advantages

- » Most likely to offer useful search, navigation and visualisation functionality;
- » Reach a wider audience of potential users;
- » Accepts a wide range of data types;
- » Suitable for cross-disciplinary data.

Disadvantages

- » Requires scrutiny of terms and conditions to ensure consistency with your funder, journal or institution's policies on cost recovery, copyright/IP, and long-term preservation;
- » No editorial control over quality of deposited materials;
- » In most cases, only simple metadata is available, which is usually not enough for reuse.

Domain specific data repository

Advantages

- » Offers specialist domain knowledge and data management expertise, e.g. to create a catalogue record and documentation;
- » Likely to accept complete datasets (and not only the part of the dataset on which a publication is based);
- » May make your data visible via dissemination and promotion.

Disadvantages

- » Likely to be selective about what kind of data they accept.

Trusted domain specific data repository

Advantages

- » Offers specialist domain knowledge and data management expertise, e.g. to create a catalogue record and documentation;
- » More likely to accept complete datasets;
- » Provides preservation and curation to community standards, e.g. file formats migration;
- » Ability to control access of (sensitive) personal data;
- » May handle data re-use queries;
- » May make your data visible via dissemination and promotion.

Disadvantages

- » Most likely to be selective about what kind of data they accept;
- » May charge for data publishing;
- » Requires advance planning of the effort needed to meet high standards for metadata and documentation.

Choosing a data repository

There are hundreds of repositories worldwide. Some cater a specific research domain, while others are general-purpose repositories. They may be called something other than a repository, for example, a data centre or an archive | Whyte (2015).

If you decide to choose a data repository for publishing your data, which data repository should you choose? Sometimes the repository is already determined by your funder or another external party. But if the choice is yours to make, you may consider following the order of preference in the [recommendations by OpenAIRE](#) (2016b):

» **1: A (trusted) domain repository**

Use a (trusted) repository already established for your research domain. The [CESSDA archives](#) are examples of domain-specific trusted repositories. Do note that not all individual datasets may be accepted or only certain types of data (e.g. surveys but not qualitative data). As a general rule, high-quality data with a potential for reuse and that can be publicly shared are submitted to this kind of repositories.

» **2: An institutional or recommended data repository**

If a domain repository is not available, use an institutional research data repository. If such a repository is not available, you may follow the guidelines of your university or publisher. Some publishers provide lists of recommendations e.g., PLoS ONE (2014b) [recommended repositories](#).

» **3: A general purpose repository**

If none of the above is available, use a general purpose repository like [Zenodo](#) (n.d.), [Figshare](#) (n.d.) or [Harvard Dataverse](#) (2017). Here you can store, share and register your research data. Do take note that long-term preservation of your data collection is not always guaranteed. Check the repository in question to find out.

» **4: Find your own at re3data.org**

Search [Re3data.org](#) (n.d.), a registry of over 1500 research data repositories, to discover other data repositories. You can search by subject, content type, and country. In addition, you can select whether you want to search for data archives with a certificate (a trusted repository), with data sets that are available via open access or for data sets that have a persistent identifier.

Expert tips



Timing is everything!

In data archiving and publishing timing is everything. If you archive or publish your data as soon as data collection ends, your knowledge about your data is still very high. As such, it will take you the least time to prepare your data for deposit while simultaneously guaranteeing the highest possible data quality for future users.

Publish a data paper

For high-quality datasets consider publishing a data paper in a data journal. This way, you can describe your datasets in more detail, which will increase their visibility and chances of being re-used. The data journal does not hold the datasets (they are in a data repository). See [‘Promoting your data’](#) for more information on this route.

Choose between self-archiving and expert help

There is a difference between self-archiving without any help and archiving with the help of an expert. While self-archiving is a quick and easy way to publish data, archiving with the help of an expert will enhance data quality. Expert help is most likely to be available at a trusted domain repository and an institutional repository. Check to see whether that is the case.

6.4 Publishing with CESSDA archives

For high-quality data with a potential for reuse, we recommend you to assure long-term access by publishing your data with a trusted repository, like many of the CESSDA archives. CESSDA archives aim to make the research data accessible with as few restrictions as possible, while at the same time protecting (sensitive) personal data from inappropriate access.

CESSDA archives per country

If you decide to publish your data to one of the CESSDA archives you will have to invest some time and effort to prepare your data. If research data management is a vital part of your work, then the majority of work has already been done on your way.

See the CESSDA website for an overview of national CESSDA archives:

<https://www.cessda.eu/About/Consortium>

Added benefits of a CESSDA repository

As opposed to self-archiving your dataset, publishing your dataset at a CESSDA archive has a great advantage of having expert help within reach. CESSDA research data management experts can help you to increase the comprehensibility, visibility, findability, reusability, longevity and the overall quality of your datasets in numerous ways.

For a list of benefits and requirements, consult the online version of this guide:

<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives>

Do you want to dive in deeper?

For data licensing, data citation and data access we have prepared additional information in the following subchapters.

6.4.1 Citing your data

Persistent identifiers ensure future access to unique published digital objects, such as a text or data set. Persistent identifiers are assigned to data sets by digital archives | American Sociological Review - [Submission Guidelines](#) (Sage Publishing, 2017).

For data products to be uniquely identifiable and attributable to their data creators two types of identifiers are recommended:

» **A persistent identifier (PID) to your dataset**

The publication of data sets is becoming more and more important as a citable contribution to research. To become citable, you need to make sure that your datasets gets a unique, persistent identifier. The Digital Object Identifier (DOI) is a well-known identifier in academia. Having a PID is an important aspect of making sure your data meets the F (Findability) and A (Accessibility) in [FAIR data management](#).

» **A persistent author identifier**

To make your research results even more connected you can create your personal persistent author identifier. The [ORCID iD](#) provides such a persistent digital identifier, distinguishing you from every other contributor and supporting automated linkages among all your professional activities. By creating and using an ORCID iD you will be able to present all of your - growing - work through one channel.

Citing new data types

Citing rapidly changing data is also challenging. The Data Cite organization has published a [recommendation](#) regarding citing new data types. There is the possibility to cite the continuously updated dataset and only add an access date and time to the citation. However, this means that the citation does not result in access to the resource as cited when it was changed in the meantime. This limits reproducibility of the work that uses this form of citation. Another option is to cite a specific "snapshot" (i.e., a copy of the entire dataset made at a specific time) but this requires unique identifiers for each version/snapshot of data.

Data Citation and impact

Data citation is the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to other scholarly resources | [Australian National Data Service](#) (n.d).

The impact of your research may be determined by a wide range of research outputs such as data sets, software, blog posts, presentations, tweets, etc. Being able to cite such research outputs is important for building a culture where all types of research outputs count. In the video (Research Data Netherlands, 2014) linked below data citation and the role of persistent identifiers is explained:

<https://www.youtube.com/watch?v=PgqtiY7oZ6k>

Expert tips



1. Deposit your data in a data repository

When you deposit your data in a (trusted) data repository, a persistent identifier to your data sets is often automatically assigned.

2. Register for an ORCID iD

Registering for an ORCID iD is easy. [Do it now](#) (ORCID, n.d.)! Or first have a look at [this video](#) (Vanhaverbeke, 2017) in which other researchers state how having an ORCID iD benefits them.

3. Check how FAIR your data are

Want to know how FAIR your data are? Have a look at the [checklist by Jones and Grootveld](#) (2017).

4. Include persistent identifiers as a variable

Include the persistent identifier to your dataset as a variable in your data file. For example, [the database from the ISSP 2015 on Work Orientations](#) (GESIS, n.d.) includes the following variable: name of the variable: DOI; variable label: "Digital Object Identifier". It has the same value for all the cases: [doi:10.4232/1.12848](#). The link goes directly to the metadata in the [GESIS data archive](#).

6.4.2 Licensing your data

If you publish your data in a data repository of your choice, a licence agreement will be applied to your data. A licence agreement is a legal arrangement between the creator/depositor of the data set and the data repository, signifying what a user is allowed to do with the data. Stating clear re-use rights is like having a warm 'Welcome' on the doormat of your dataset. It is an important aspect of making sure your data meet the R (Reusable) in [FAIR data management](#).

To make re-use as likely as possible we advise you to choose a licence which:

- » Makes data available to the widest audience possible;
- » Makes the widest range of uses possible.

About Creative Commons licences

The main attributes of using [Creative Commons \(2017\)](#) licences for the licensing of data, datasets, and databases (Korn and Oppenheim, 2011) are:

- » The ease of use of the licences;
- » The widespread adoption of the licences;
- » Their flexibility;
- » Their availability in human-readable and machine-readable forms allowing both researchers and computers to immediately know what they are allowed to do with your data;
- » The chance that your data are reused.

There are 7 licences for which the details are given in the table below (inspired by Foter, 2015):

Licence	Can I copy & redistribute the work?	Is it required to attribute the author?	Can I use the work commercially?	Am I allowed to adapt the work?	Can I change the licence when redistributing?
CC0	Y	N	Y	Y	Y
CC BY	Y	Y	Y	Y	Y
CC BY-SA	Y	Y	Y	Y	N
CC BY-ND	Y	Y	Y	N	Y
CC BY-NC	Y	Y	N	Y	Y
CC BY-NC-SA	Y	Y	N	Y	N
CC BY-NC-ND	Y	Y	N	N	Y

Do note that a CC licence cannot be revoked once it has been issued.

The licence you are allowed to apply may be determined or limited by the data repository of your choice. An example is given in the following box.

Data licences at the Slovenian Social Science Data Archives

The [Slovenian Social Science Data Archives](#) (ADP, 2017b) allows you to choose between [three types of Creative Commons licenses](#) (ADP, 2017c):

- » CC-BY
Users:
 - » Are free to share — copy and redistribute the material in any medium or format;
 - » Are free to adapt — remix, transform, and build upon the material;
 - » May use the data sets for any purpose, even commercially.
- » CC-BY-NC
Users:
 - » Are free to share — copy and redistribute the material in any medium or format;
 - » Are free to adapt — remix, transform, and build upon the material;
 - » May not use the data sets for commercial purposes.

Both licenses have the condition of Attribution. A user must give [appropriate credit](#) (Creative Commons, n.d.a).

Recently, the ADP also gives the possibility to choose a Creative Commons Zero License (in short: [CC0](#) or CC Zero Waiver (Creative Commons, n.d.b.)). With this licence, the depositor waives all rights to the data.

Considerations in choosing a licence

If you only consider your own benefit, you might choose a licence for which attribution is required. What you may not realise is that when such data is blended with similarly licensed data [this may lead to impracticalities of required attribution](#) (Dodds, 2014) whenever the data is reused. To facilitate the release of datasets and databases into the public domain, Creative Commons created the CC0 licence.

CC0 is the only truly open Creative Commons licence. The copyright owner waives all its rights, including the database right and the right to be identified as the creator.

Although CC0 can be used to prevent attribution stacking, attribution can be important as a means of recognising both the source and the authority of the data. To acknowledge this right, the use of CC0 can include the publishing of non-binding suggestions for best practices in attribution.

There will be circumstances in which CC0 is inappropriate, due to specific risks that might arise for the licensor and perhaps subsequently also for any users. E.g. when:

- » Datasets containing (sensitive) personal information are deposited for which consent has not been cleared (see the [chapter on protecting data](#));
- » Permission of the copyright holder has not been sought;
- » The rights holders are unknown or cannot be traced (orphan works).

In these cases, licences that place 'some' restrictions upon the user, such as those with an "ND" (No derivatives) and/or "NC" (Non-Commercial) might be more appropriate.

Tips for choosing a licence

1. Be sure who owns the data

Remember you can only archive and publish data you own (or if you have permission).

2. Use the licence selector

Choose an appropriate licence for your datasets with [this licence selector](#) (n.d.).

6.4.3 Access categories

Publishing data in a data repository does not automatically make them openly accessible. (Sensitive) personal data can still be protected by limiting access to the data. Access controls can permit control down to an individual file level, meaning that mixed levels of access control can be applied to a data collection.

Many data repositories operate a three-tiered approach to data access:

» **Open access**

Data that can be accessed by any user whether they are registered or not. Data in this category should not contain personal information unless consent is given (see '[Informed consent](#)').

» **Access for registered users (safeguarded)**

Data that is accessible only to users who have registered with the archive. This data contains no direct identifiers but there may be a risk of disclosure through the linking of indirect identifiers.

» **Restricted access**

Access is limited and can only be granted upon request. This access category is for the most sensitive data that may contain disclosive information. Restricted access requires the long-term commitment of the researcher or person responsible for the data to handle the upcoming permission requests.

» **Embargo**

Besides offering the opportunity for restricted access 'for eternity' most data repositories allow you to place a temporary embargo on your data. During the embargo period, only the description of the dataset is published. The data themselves will become available in open access after a certain period of time.

Access conditions may differ slightly between data repositories. In the boxes below, two examples are given.

Access regulation of the Slovenian Social Science Data Archives (ADP)

At the Slovenian Social Science Data Archives (ADP) access to data and accompanying materials is determined in the [Policy of Digital Preservation](#) (ADP, 2017b). The types of access in ADP are the following:

» **Open Access**

Users may freely access the catalogue of the ADP, study metadata and research data of a limited range of studies without registration. Nonetheless, the use of data and accompanying materials is limited by the legislation, the social sciences and institutional ethical standards and copyright.

» **Standard Access**

Standard access includes the possibility to access most of the research data in the ADP. In order to obtain standard access, users need to fill in the Registration form to access materials in the Catalogue of the ADP. The users need to identify, define the terms of use of research data and comply with the General Provisions of and Terms of Use of the ADP. Research data that may be accessed by a standard registered user, are fully anonymized. These are the so-called Public Use Files (PUF).

» **Special Conditions Access**

Some data sets are only accessible under special conditions. In order to gain access, a special permission from the original authors is needed. For example when:

» **Data are not fully anonymized**

In this case, additional protection is required. Such files are called Scientific Use Files (SUF).

» Embargo

The authors place an embargo on access and decide that the datasets will only be available after a certain period of time, for example, after 6 months.

» Limited availability

The dataset is available only to the person/institution, ordering the study, or to the original authors.

If a user wants access to files in the Special Conditions Access section, not only a regular registration form (Standard Access) should be filled in but also an additional one which is called: "Application for access to materials on request". The Commission for the protection of Confidentiality carefully inspects such applications and decides on the possibility of access to the requested study data.

Types of possible special access are:

» Access through a safe connection**» Access in a safe environment**

If the data are especially sensitive the user may be granted access to it only in a safe room of the data archive. These are the so-called Secure Use Files (ScUF) that the user may access only after signing a special contract, determining the rights and obligations of use of the requested research data.

Access regulation at DANS, the Netherlands

All research data at [DANS](#) are stored in and made available by its online repository [EASY](#) (DANS, 2017b). A licence agreement is always agreed between DANS and the depositor of the dataset: the person or organisation depositing a dataset in EASY who is normally the rights holder. One of the most important parts of this licence agreement is the access category by which the access to the dataset can be specified.

DANS supports the Open Access movement. This means that DANS encourages research data and publications to be made freely available as much as possible, without any restrictions. However, substantiated reasons exist why research data is not, or not immediately, freely accessible. This can be due to the presence of personal data or a temporary embargo on data due to an impending PhD thesis or other publication, contract obligations with third parties, etc. DANS, therefore, provides along with open access, the possibility of restricted access to research data.

EASY offers two Open Access categories and one Restricted Access category. The access categories are:

» Open Access ([CC0 Waiver](#), Creative Commons, n.d.b.)

The dataset is, without any restriction, made available to all EASY users, both registered and unregistered, in accordance with the conditions of the Creative Commons Zero Waiver.

» Open Access for Registered Users

The dataset is only made available to all registered EASY users. Any existing copyrights and/or database rights are respected.

» Restricted Access

The dataset is only made available to those registered users that have obtained permission from the rights holder.

Datasets containing personal data are mostly placed in the category of Restricted Access. Some datasets with personal data are made available in the Open Access categories. This is, however, only possible when explicit informed consent has been given by the persons involved. This is quite often the case with [Oral History interviews](#) (DANS, 2012). Besides from this open category, sensitive data can only be accessed by authorised users whose identities have been checked and who may be required to also sign

special, additional, conditions of use.

Open metadata for (sensitive) personal data

Even if personal data cannot be published in open access, it is always possible to publish the metadata that belong to this dataset. Openly publishing metadata is, in fact, the only way to make such datasets discoverable.

Trusted data repositories are dedicated to increasing the discoverability of your data sets. Therefore, metadata are always freely accessible in any of the CESSDA archives. That means that:

1. No registration is needed for searching in the metadata;
2. No registration is needed for harvesting the metadata (e.g. by search engines).

Metadata of sensitive datasets should never contain confidential or identifying elements or characteristics, like names.

When someone finds a dataset under restricted access (most likely because they contain (sensitive) personal data), he or she can submit an access request to the rights holder. If this is granted, the dataset will be available for download by this user. Even then the use is restricted. The user is not allowed to make the personal data of this data set public and can only refer to the data in an anonymised way.

Access control strategy

When choosing an access category, consider the following:

- » Does the data contain identifiable information?
- » Can the information in this data collection be linked with anything in another data collection which might lead to participant's identities being disclosed?
- » What did participants consent to?
- » If 'restricted access' is to be chosen who will manage the access to this request?

6.5 Promoting your data

How can you attract people to use your data and make them as impactful as possible? Consider promoting reuse of your data in one of the following ways:

Choose open access

If you deposit your data in a data repository, choose open access. If researchers can easily access your data, it is more likely that your data will be re-used and have an impact on their work.

License your data

Licensing your data is a prerequisite for data impact. If researchers are unclear about what they are allowed to do with your data, they might not use it at all (see [‘Licensing your data’](#)).

Always cite your data

Always cite your data and link your data to scientific publications which are based on this data.

How to cite data: an example

The following dataset which holds data on studying migrations patterns in the Summer Olympics between 1948 and 2012 covers approximately 40,000 athletes and contains information on the country they represented as well as their country of birth. According to the data repository, which holds this open access dataset, it should be cited as:

Reference: Jansen, J. (Erasmus University Rotterdam) (2017): Foreign-born Olympic athletes 1948 - 2012. DANS. <https://doi.org/10.17026/dans-2xf-pyqp>

Also, see the [data citation paragraph](#) in this tour guide.

Publish in a data journal

Consider publishing an article in a peer-reviewed data journal. Data journals are designed to comprehensively document and publish deposited datasets and to facilitate their online exploration. Recommendations for such journals for social sciences and humanities are:

Research Data Journal for the Humanities and Social Sciences ([RDJ](#), Brill, 2017);

Journal of Open Psychology Data ([JOPD](#), Ubiquity Press, n.d.a);

Journal of Open Archaeology Data ([JOAD](#), Ubiquity Press, n.d.b);

[Open Health Data](#) (Ubiquity Press, n.d.c.).

Tip Read this blogpost: [Introducing the ‘data paper’ in the Research Data Journal for the Humanities and Social Sciences](#) (Moody, 2017).

Teach with your dataset

Consider preparing a lecture using your datasets (or that of others) and prepare video tutorials on how to use the dataset.

Datasets for training purposes: easySHARE

The Survey of Health, Ageing, and Retirement in Europe ([SHARE-ERIC](#), 2017a) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks. Surveys are organised bi-annually since 2004. SHARE currently covers [27 European countries and Israel](#) (SHARE-ERIC, 2017b).

The SHARE database is easily accessible to the entire research community; data from the SHARE Waves 1 to 6 are available since 2017. A longitudinal data set “[easySHARE](#)” (SHARE-ERIC, 2017c) has been

created especially for training purposes. It contains only selected variables merged into a single data file. It is more user-friendly than the complete set of SHARE panel data.

Choose a data repository which promotes your data

You can promote your own data. In addition, you can choose a data repository which promotes data for you.

For examples of how CESSDA archives promote your data, see the online version of this guide:

<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Promoting-your-data>

Grow your data's impact with altmetrics

Altmetrics, or 'alternate metrics', are alternative parameters which measure the impact of your research. More and more, research data and software code are shared in data repositories and quoted in publications. More and more repositories attach a DOI - a persistent identifier - to such datasets, allowing to count how often a dataset:

- » Has been cited;
- » Has been viewed or downloaded;
- » Has been stored in online literature management systems;
- » Is listed in online news media or social media.

After you have uploaded your dataset to a data repository which adds DOIs, consider to:

- » Write a blog post or an article about your data publication;
- » Tweet about it;
- » Write about it on Facebook.
- » Etc.

Do not forget to always cite your dataset, everywhere (in your publications, in blog posts, in social media, etc.). It is the only way to keep track of how your data is used, viewed, liked. See 'Data citation' for information on how to cite your data.

Tracking data publications

Data publications can be tracked by (Ball and Duke, 2015):

Citation-based metrics

- » [DataSearch](#) (Elsevier, 2017)
Searches data repositories, including figures/tables and has a preview option so you can judge whether the data are useful.
- » [DataCite](#) (n.d.)
Searches datasets that have been given a DOI.
- » [Data Citation Index](#) (Clarivate Analytics, 2017a). (Web of Science - licensed database with paid access only)
Searches the metadata of the datasets for [connected data repositories](#) (Clarivate Analytics, 2017b).

Altmetrics-based metrics

- » [ImpactStory](#) (n.d.)
- » [PlumX](#) (Plum Analytics, 2015)
- » [Altmetric](#) (n.d.)
- » Data repositories (Downloads and views counts)

6.6 Adapt your DMP: part 6



This is the sixth 'Adapt your DMP' section in this tour guide. To adapt your DMP, consider the following elements and corresponding questions:

Deposit your data

- » Will the data you produce and/or used in the project be usable by third parties, in particular after the end of the project?
- » Which data and associated metadata, documentation and code will be deposited?
- » What methods or software tools are needed to access the data?
- » Is documentation about the software needed to access the data included?
- » Is it possible to include the relevant software (e.g. in open source code)?
- » What data quality assurance processes will you apply?
- » Will the application of a persistent identifier to your data be ensured?

Deposit timing and duration

- » When will your data be made available for re-use? Is there an embargo period?
- » How long does the data need to be retained? For how long should the data remain reusable?

Access category

- » How will the data be made available? What access category will you choose?
- » When thinking about access categories consider the following:
 - » What did the participants consent to?
 - » Does the data contain anything sensitive?
 - » Can the information in this data collection be linked with anything in another data collection which might lead to participant's identities being disclosed?
 - » If 'restricted access' is to be chosen who will manage the access to this request?

Data licensing

- » How will your data be licensed to permit the widest re-use possible?
- » Have you considered which kind of licence is appropriate for sharing your data and what, if any, restrictions there might be on re-use?
- » If you are purchasing or re-using someone else's data sources have you considered how that data might be shareable, for example negotiating a new licence with the original supplier?

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

[Please see the online version of this guide.](#)

Chapter 7

Discover

Contents

Main take-aways	177
7.1 The process of data discovery	178
7.2 Data repositories as data resources	193
7.3 Resources for social media data.....	198
7.4 Access, use and cite data	200
7.5 Adapt your DMP: part 7	206
Sources and further reading	207

[View the online version of this chapter](#)

Main authors of this chapter

Johana Chylikova, Czech Social Science Data Archive (CSDA)

Martin Vávra, Czech Social Science Data Archive (CSDA)

Jindrich Krejčí, Czech Social Science Data Archive (CSDA)

Jennifer Buckley, UK Data Service, University of Manchester

Michaela Kudrnacova, Czech Social Science Data Archive (CSDA)

Introduction



If you want to reuse or review research data shared by other researchers, this chapter is for you. We will show you the steps you can take in your process of data discovery, from developing a clear picture of the data you need to evaluating data quality.

To make it easier for you to discover high-quality data, we present curated lists of different types of social science data sources in Europe and around the world. The chapter concludes with things to keep in mind when you access selected data.

Main take-aways

After completing your travels through this chapter on data discovery you should:

- » Be able to set up - and adjust - a search strategy to find suitable data for your research purposes;
- » Understand that social science data repositories are important sources for discovering social science data;
- » Be aware of data sources which CESSDA-experts recommend for selected research topics;
- » Be aware of steps in evaluating the quality and usefulness of data for secondary analysis;
- » Understand different types and modes of access to data;
- » Be able to answer the [DMP questions](#) which are listed at the end of this chapter, and adapt your own DMP.

7.1 The process of data discovery

A fictive data discovery story with roots in reality

Jana Svoboda is an economist and works in a public research institute in the Czech Republic. She needs international comparative data on work orientations. How did she discover and access such data?

Find out how

In preparation of her research project, Jana takes the following steps to find relevant data:

» **She reviews the literature on the topic**

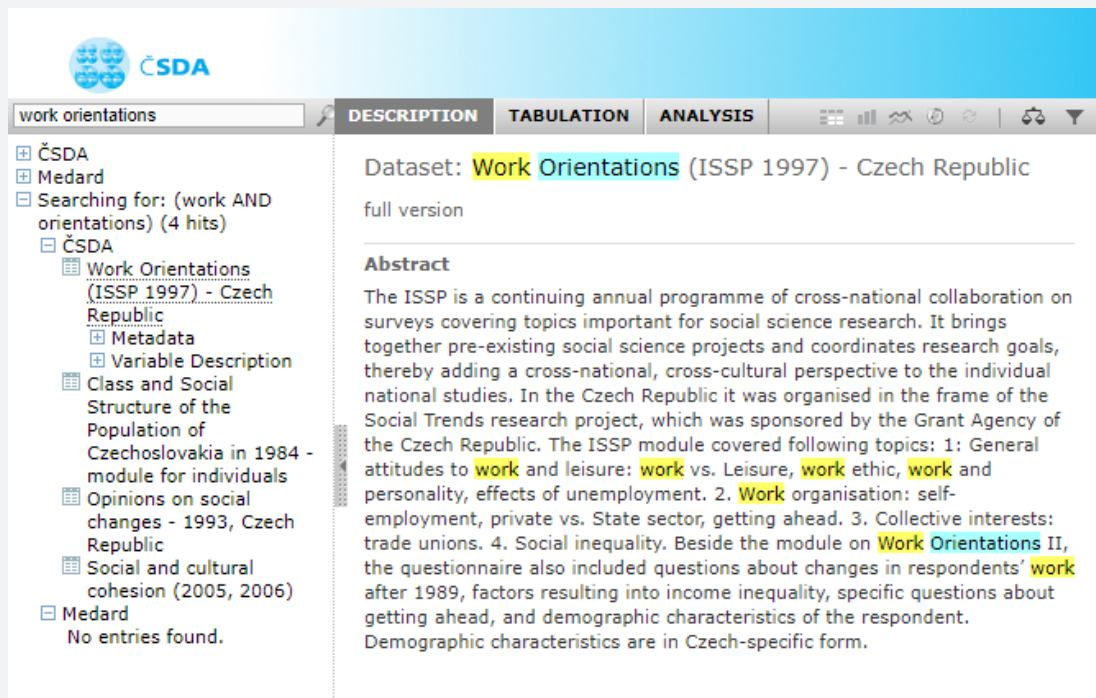
Jana begins with a review of scholarly literature and looks for data used by other researchers. She has read a number of studies on the topic before and now she reviews the ‘research method’ and ‘data’ sections in articles to learn about the data and the data resources. The authors mention several publicly available datasets, that may be used in Jana’s research. The study of the International Social Survey Programme (ISSP) on Work Orientations is among them. However, only a few datasets are properly cited and the persistent links (DOI) are missing for most of them. She must look elsewhere to find out details about the ISSP Work Orientations studies.

» **She looks for data at the survey programme website**

Jana visits the [ISSP website](#) (ISSP, n.d.). There she finds general information about the Work Orientation surveys, methodology and participating countries. She downloads the international module questionnaire, i.e. the source questionnaire written in English whose translation was used in individual countries. This questionnaire contains all variables measured in all ISSP countries in 2015. Jana reads the questionnaire and finds out that it contains variables that she might use in her study. She follows the link to the [ISSP data archive at GESIS](#) (GESIS, n.d.a). The archive provides rich metadata from each ISSP survey and enables users to download data for scientific research and teaching purposes. The [GESIS ZACAT](#) data catalogue also offers its users the ability to do a very simple analysis right in the online environment. Jana browses the variables in individual Work Orientation studies in the online archive and finds important variables for testing her research hypothesis. Then she downloads datasets from several Work Orientation surveys that were conducted in many countries in 1989, 1997, 2005 and 2015.

» **She searches for data in social science data archives**

Jana searches the GESIS data catalogue also for other data on work orientations. Besides the ISSP, there are over a hundred other studies. However, they are not comparative by nature or the topic of work orientations is not so central. Jana decides to visit the [Czech Social Science Data Archive](#) (CSDA, n.d.) to look for data that give her a more detailed view on work orientations in her country. She finds out that ISSP Work Orientation datasets which only contain data from the Czech Republic, include variables that were not part of the international “core” questionnaire and were measured only in Czechia. These country specific variables allow Jana a more detailed and deeper analysis of work orientations.



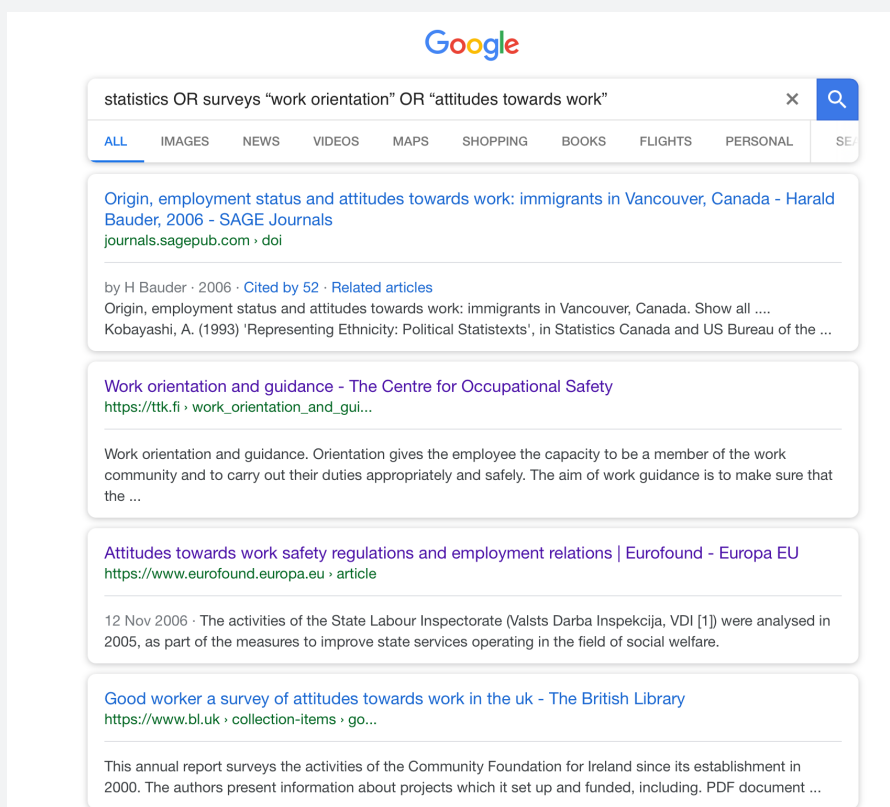
The screenshot shows the CSDA (CESSDA Data Management Expert Guide) interface. The search term 'work orientations' is entered in the search bar. The left sidebar shows a tree view of search results, with 'Work Orientations (ISSP 1997) - Czech Republic' selected. The main content area displays the dataset title, a description, and an abstract. The abstract text is highlighted in yellow.

Dataset: Work Orientations (ISSP 1997) - Czech Republic
full version

Abstract
The ISSP is a continuing annual programme of cross-national collaboration on surveys covering topics important for social science research. It brings together pre-existing social science projects and coordinates research goals, thereby adding a cross-national, cross-cultural perspective to the individual national studies. In the Czech Republic it was organised in the frame of the Social Trends research project, which was sponsored by the Grant Agency of the Czech Republic. The ISSP module covered following topics: 1: General attitudes to work and leisure: work vs. Leisure, work ethic, work and personality, effects of unemployment. 2. Work organisation: self-employment, private vs. State sector, getting ahead. 3. Collective interests: trade unions. 4. Social inequality. Beside the module on Work Orientations II, the questionnaire also included questions about changes in respondents' work after 1989, factors resulting into income inequality, specific questions about getting ahead, and demographic characteristics of the respondent. Demographic characteristics are in Czech-specific form.

» **She searches for other data on the web**

Jana wants a complete picture of the available data, so she continues searching. She uses Google and employs various keywords such as statistics, surveys or questionnaires and combines them with her research topic (work orientation OR attitudes towards work OR labour force). Jana learns about a few interesting organisations which host datasets on work orientations such as the [European Working Conditions Surveys \(EWCS\)](#), one of the [datasets maintained by Eurofond](#) (Eurofond, n.d.).



The screenshot shows a Google search results page. The search query is 'statistics OR surveys "work orientation" OR "attitudes towards work"'. The search results are displayed in a list format, with the first result being 'Origin, employment status and attitudes towards work: immigrants in Vancouver, Canada - Harald Bauder, 2006 - SAGE Journals'. Other results include 'Work orientation and guidance - The Centre for Occupational Safety', 'Attitudes towards work safety regulations and employment relations | Eurofound - Europa EU', and 'Good worker a survey of attitudes towards work in the uk - The British Library'.

Google

statistics OR surveys "work orientation" OR "attitudes towards work"

ALL IMAGES NEWS VIDEOS MAPS SHOPPING BOOKS FLIGHTS PERSONAL

Origin, employment status and attitudes towards work: immigrants in Vancouver, Canada - Harald Bauder, 2006 - SAGE Journals
journals.sagepub.com · doi

by H Bauder · 2006 · Cited by 52 · Related articles
Origin, employment status and attitudes towards work: immigrants in Vancouver, Canada. Show all ...
Kobayashi, A. (1993) 'Representing Ethnicity: Political Statistexts', in Statistics Canada and US Bureau of the ...

Work orientation and guidance - The Centre for Occupational Safety
https://ttk.fi · work_orientation_and_gui...

Work orientation and guidance. Orientation gives the employee the capacity to be a member of the work community and to carry out their duties appropriately and safely. The aim of work guidance is to make sure that the ...

Attitudes towards work safety regulations and employment relations | Eurofound - Europa EU
https://www.eurofound.europa.eu · article

12 Nov 2006 · The activities of the State Labour Inspectorate (Valsts Darba Inspekcija, VDI [1]) were analysed in 2005, as part of the measures to improve state services operating in the field of social welfare.

Good worker a survey of attitudes towards work in the uk - The British Library
https://www.bl.uk · collection-items · go...

This annual report surveys the activities of the Community Foundation for Ireland since its establishment in 2000. The authors present information about projects which it set up and funded, including. PDF document ...

Each empirical research project should start by searching for existing data resources relevant to the research topic. This is essential for projects based on secondary analysis (which reuse data produced by another research project), but also important for projects that intend to collect original data.

When you discover existing data, you can use them to your advantage in the following ways:

Reuse data and save costs and time

Using existing data is a cost- and time-saving way to carry out your own research. You can use data in the same way as the researchers before you, or you can look for new perspectives and use the data differently.

Compare results or make replication studies

Adopting previously used elements of research design allows you to compare your results across time and internationally and allows you to make replication studies. Many collaborative projects such as the [International Social Survey Programme](#) (ISSP, n.d.) or the [European Social Survey](#) (ESS, n.d.) make their data publicly available and rely on a culture of data sharing and open access.

Reuse verified elements of research design

Existing databases and their metadata allow you to check the measurement instruments and other elements of study design that have been tested in prior research. You could use such verified elements of research design in your own data collection.

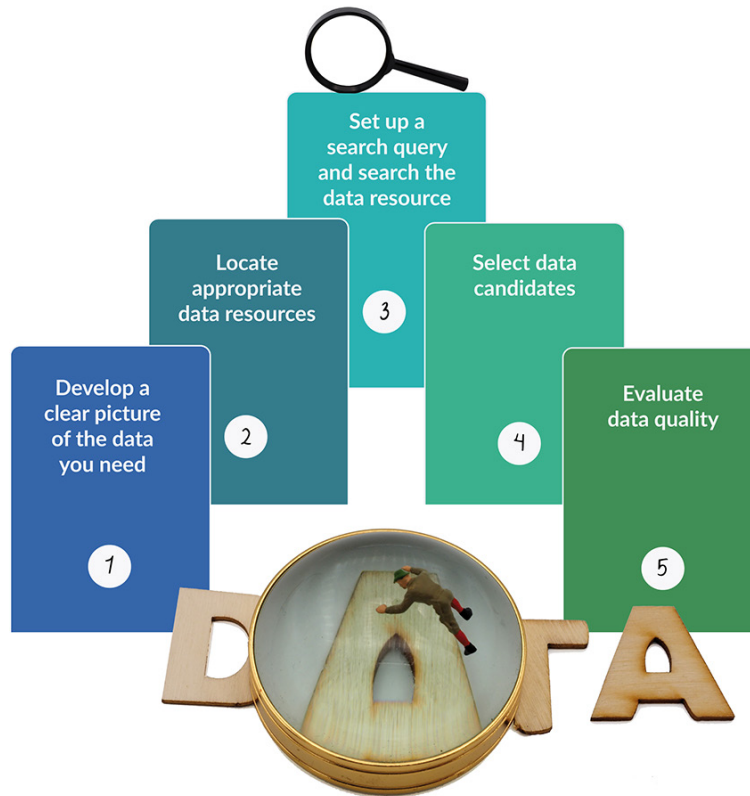
Enhance data quality and foster innovation

Discovering existing data helps you to adopt existing research standards, embed your research into a contemporary state of knowledge and make your study more innovative.

A disadvantage of using existing data may be that the research design is set and you must be satisfied with the exact wording of questionnaire items, population, sampling etc. Moreover, you can not influence the quality of data (Gregory, et al., 2018a). Therefore, the quality of the metadata must be as high as possible, so that you have sufficient information to decide whether or not you want to use the data.

Steps in data discovery

Data discovery is a process of several distinct - and cyclic - steps. You can structure your search according to the following steps (inspired by Gregory, et al., 2018b):



1. Develop a clear picture of the research data you need

In the process of data discovery, it's important to be aware of the type of data you are looking for. What data fit your research intentions?

The term "research data" can be broadly understood as any data usable in research, or more narrowly, as data produced for research purposes. In this guide, we focus on data generated in social science research. Some data sources specialise in certain types of data. Specifically, there are different methodologies and different types of data used in quantitative and qualitative social science research. Please read Chapter 1 for an introduction to the [different types of research data](#) and [concepts of quantitative and qualitative data](#).

Listing the characteristics of the data you want to discover makes it easier:

- » to formulate the right search terms to find sources which hold such data;
- » to search the data source of choice for adequate data.

To develop a clear picture of the research data you want to discover and use, ask yourself:

I. What is the theme/domain you study?

Before you start looking for data, you must be sure which theme or domain you are interested in. It may be politics, health, family, social inequalities etc.

II. What is your research question?

Before you start looking for right data for you, you must be sure about your research intentions. What is your research question? A research question is one or more questions that your study wants to answer. For example: “Do reading habits in childhood relate to attained education?” or “Are anti-immigrant sentiments related to age and education?”

III. What are the constructs you want to work with?

Your research question contains several constructs, i.e. scientific concepts developed for systematic inquiry of the issue. Examples of such constructs are “employment”, “attitudes to immigration”, “age” or “education”. When you are looking for appropriate data for your research, look out for indicators of such constructs. In survey research, these indicators are the variables contained in the dataset. The construct of “education” may have various indicators, the most common being “highest completed education” measured in (standardised) categories, or “years of schooling” measured in total years spent by a respondent in schools.

There also exist complex concepts that use information from more than one survey question/indicator. For example, political participation is a multidimensional concept involving voting, organisation membership and demonstrating etc.

IV. How will you operationalise the constructs?

What indicators of constructs do you need to find in the data? Do you have a preferred level of measurement for your key variables, i.e. are you looking for variables measured on nominal, ordinal or interval level?

V. What is your theory?

Does your research follow a previously developed theory? It definitely should. Look for concepts and their indicators that were previously used by other researchers to answer your research questions.

VI. What study will you perform?

E.g. you may want to use the data for a:

New original study

Examples:

- » You will use one or multiple data sources, you may combine micro (individuals) and macro (countries) level;
- » You may use only secondary data or you may combine secondary data with primary data, i.e. the data you collected within your project (“your data”);
- » You want to use the design/methodology from some other study;
- » You want to use some features of study/questionnaire in your own study (interview schedules, measurement instruments, sampling strategies etc.).

Replication study

You are repeating/replicating a study that was carried out earlier by you or somebody else.

Teaching purposes

You want to use data in teaching, perhaps datasets that were made specifically for training purposes, such as [easySHARE](#) (SHARE-ERIC, 2018) or [European Social Survey Education Net](#) (ESS EduNet, 2016).

VI. What study will you perform?

E.g. you may want to use the data for a:

New original study

Examples:

- » You will use one or multiple data sources, you may combine micro (individuals) and macro (countries) level;
- » You may use only secondary data or you may combine secondary data with primary data, i.e. the data you collected within your project ("your data");
- » You want to use the design/methodology from some other study;
- » You want to use some features of study/questionnaire in your own study (interview schedules, measurement instruments, sampling strategies etc.).

Replication study

You are repeating/replicating a study that was carried out earlier by you or somebody else.

Teaching purposes

You want to use data in teaching, perhaps datasets that were made specifically for training purposes, such as easySHARE (SHARE-ERIC, 2018) or European Social Survey Education Net (ESS EduNet, 2016).

VII. What specific characteristics should the data have?

First make a detailed plan of your study and specify its key characteristics (Babbie, 1998):

Are you going to use quantitative or qualitative approach?

- » Will you use survey data (= answers of respondents on standardized questions) or other types of quantitative data (e.g. administrative data, social media data etc.)?
- » Secondary analysis of qualitative data is less common and qualitative data are less available from data repositories; however, opportunities exist to reuse qualitative research outputs in a new analysis and also when designing new research projects.

What is the population you want to study?

- » Who is your target population? Adults, children, country citizens, migrants, local authorities, single mothers etc.
- » What is the unit of analysis? Individuals, households, regions, countries etc.
- » Do you need a large representative sample? If you need data to be representative of a specific population, you most likely need to find data from a sample taken using random sampling techniques. To use data from a quota sample is also possible, such a data are representative for the population in characteristics as gender, age, education etc.
- » What geographical area do you want to cover? A specific country, a specific region, all EU countries etc.

What should be the geographical origin of the data?

- » Do you want data from one country or are you going to use data from different countries?
- » Do you need national data (concerning the population of only one country)?
- » Do you need international data (concerning the population of several countries, where the methodology is the same in each country and the data are comparable across countries)?
- » What is the desired time scope?

What is the desired time scope?

E.g:

- » As recent as possible;
- » Data from a precise point in time (e.g. 2008);
- » Data from several specific time points (e.g. 2009 and 2014);
- » Longitudinal data covering a time span to track differences in time;
- » Cross sectional data (to analyse data from a population, or a representative subset, at a specific point in time);
- » Longitudinal (or panel) data (where the same respondents are asked repeatedly (in two or more data collection waves));
- » Repeated cross sectional data (where the survey design is repeated on a different sample of respondents; respondents in the "Sample t" are different respondents than in the "Sample t+1").

VIII. Do you have other preconditions?

E.g.:

- » Do you need a specific file format?
- » Should the data be available right away in open access or is restricted access also an option?

2. Locate appropriate data resources

Once you have developed a clear picture of the data you need, you will need to locate appropriate data resources which may host such data.

Depending on what you already know about 'data repositories out there', you will probably proceed from one of the following points:

1. You know appropriate data resources (or know who to ask)

You are already acquainted with trusted data resources on your research topic, e.g. from:

- » colleagues close by or colleagues you have met at conferences, training events, etc.;
- » curated lists of data sources such as those in the paragraph '[Data repositories as data sources](#)' in this chapter.

2. You are not yet familiar with possible data resources (and don't know who to ask)

How do you discover such data resources if you do not know they exist? To discover data resources from scratch, consider using the following instruments:

I. A registry of data repositories

A registry of data repositories is a tool that offers researchers a searchable overview of many existing repositories for research data. E.g.:

- » Re3data.org (n.d.) is a registry of research data repositories which lists over 2000 data repositories

from all research disciplines. You can search by subject, content type and country. In addition, you can set some specific conditions, e. g. limiting the search to data repositories with a certificate (a trusted repository), which host data sets available via open access or which have a persistent identifier.

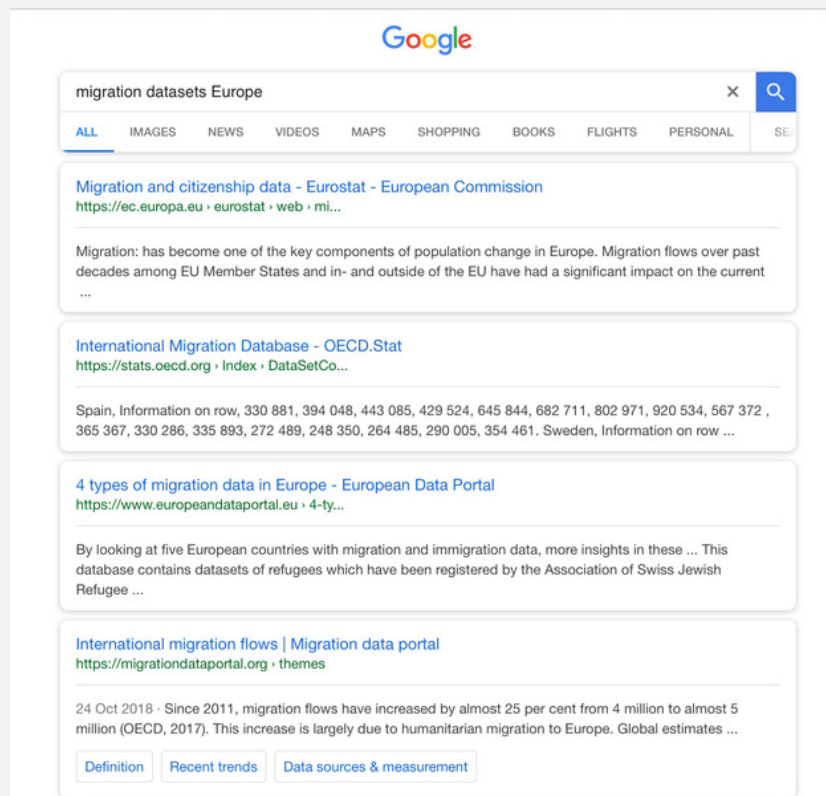
- » [OpenAIRE Explorer](#) (OpenAIRE, n.d.) provides a searchable registry of open access compatible repositories.
- » With [OpenDOAR](#) (Jisc, n.d.), the Directory of Academic Open Access Repositories, you can browse over 3500 academic open access repositories. It enables users to identify, browse and search repositories based on a range of features, such as location, software or type of material held.
- » [FAIRsharing](#) (n.d.) groups together resources (standards, databases or policies) by domain, project or organisation. It has its roots in life sciences, so the list of data repositories belonging to the domain of social sciences is not very long (December 2018).

II. A search engine or (meta)data aggregator

You can use (specialised) search engines or (meta)data aggregators for discovering relevant data sources. Examples are:

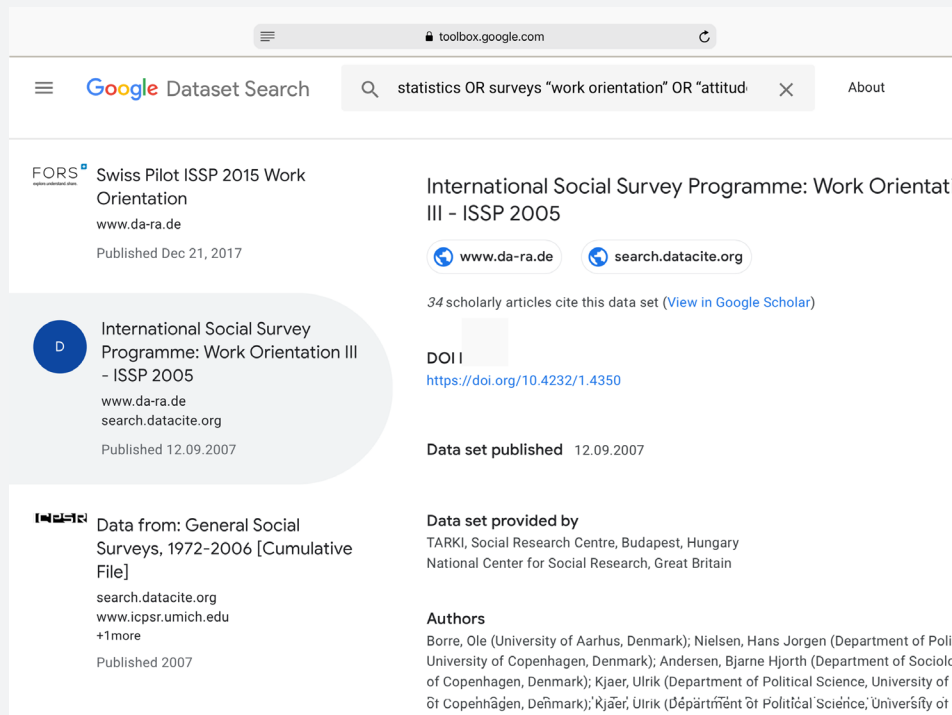
Google

You can use [Google](#) to discover organisations which hold data sets on your research topic. Apart from keywords which describe your research topic, it is important to add keywords such as 'datasets' or 'data archive' to your search query. The advantage of this approach is that you will most surely look beyond the usual suspects. Google indexes trillions of web pages. The disadvantage may be that it costs you time to filter the results.



Google Dataset Search

Google developed a tool for data search: [Google dataset search](#) (Google, n.d). The advantage is that results are already limited to data sets. The disadvantage is that you do not know what selection of data sources Google dataset search searches, so some choices have been made for you. If you do not find appropriate data it doesn't mean they do not exist. What isn't indexed, cannot be found.



The screenshot shows the Google Dataset Search interface. The search bar contains the query "statistics OR surveys "work orientation" OR "attitud". The results page displays several entries:

- Swiss Pilot ISSP 2015 Work Orientation** (FORS): www.da-ra.de, Published Dec 21, 2017.
- International Social Survey Programme: Work Orientation III - ISSP 2005**: www.da-ra.de, search.datacite.org, DOI: https://doi.org/10.4232/1.4350, Data set published 12.09.2007.
- Data from: General Social Surveys, 1972-2006 [Cumulative File]** (ICPSR): search.datacite.org, www.icpsr.umich.edu, +1more, Published 2007.

Additional information for the ISSP 2005 dataset includes: "34 scholarly articles cite this data set (View in Google Scholar)", "Data set provided by TARKI, Social Research Centre, Budapest, Hungary; National Center for Social Research, Great Britain", and a list of authors including Ole Borre, Hans Jørgen Nielsen, Bjarne Hjørth, and Ulrik Kjær.

(Meta)data aggregators

» DataCite

DataCite gathers metadata for each DOI assigned to an object. The metadata is used for a large index of research data that can be queried directly to find data, obtain stats and explore connections. You can try it for yourself at <https://search.datacite.org/> (DataCite, n.d.). All the metadata is free to access and review. The disadvantage is that you will not discover datasets with a different kind of persistent identifier.

» DataSearch

The publishing company Elsevier offers [DataSearch](#) (Elsevier, n.d.) as part of its data policy. You can filter search results by type of data or data repository. DataSearch indexes both metadata and data to facilitate the matching of queries to objects described in the research. For social sciences, ICSPR is indexed as a data source. CESSDA Archives, however, are not indexed at the moment.

III. A data catalogue

Domain aggregated data catalogues index specific selections of data resources. In the European research area the most important is the [CESSDA Data Catalogue](#) (CESSDA, n.d.a) which contains the metadata of all data in the holdings of CESSDA's service providers. It enables effective access to European social science research data. The Catalogue's search engine enables filtering by topic, data collection years, country or language. The advantage is that you do not have to query every CESSDA data archive separately. The disadvantage is that you will not find data which weren't archived by CESSDA members.

For more information about important social science data archives, visit the section '[Data repositories as data sources](#)' in this Expert Guide.

IV. A data journal

You can look for data (or rather data citations) in research papers or scholarly articles. But there is another option as well. You can use specialised data journals which publish descriptions of scientifically valuable datasets, and research texts on the sharing and reuse of scientific data. Important examples are:

- » [Research Data Journal for the Humanities and Social Sciences](#) (Brill, n.d.) is published by Brill in collaboration with DANS (one of the CESSDA archives).
- » [Scientific data](#) (Nature, n.d.) is a “branch journal” to Nature. It is mostly dedicated to natural sciences data, but publishes data articles from other fields of science too.

Expert tips:



Look out for trusted data resources

If you locate an appropriate data resource, don't forget to carefully determine the authority of the party which hosts the data. Is the data resource maintained by a trustworthy organisation?

The importance of indexing

Determine which data resources your search instrument indexes. Remember: What isn't indexed, cannot be found and may need to be discovered via a different strategy.

3. Set up a search query and search the data resource

Once you find a data resource which hosts the type of data you are interested in, you should find out how to search in the data archive or repository of choice. To translate your needs into a search request, you will have to find out what search functionalities the data resource offers. Such search functionalities differ for each individual search system.

Generally it is advised to:

I. Familiarise yourself with the structure of the data resource

When you have selected a data resource, familiarise yourself with the system the repository uses for organising data. What information about the data is contained in the repository catalogue? What metadata fields are offered? Be aware that searching in the repository catalogue mostly means that you search through the metadata (not in the data itself). This means that the quality and completeness of your results depends on both the quality of metadata and your ability to formulate the appropriate search terms for finding the data you need.

You can find more about metadata in [chapter 2 of this Data Management Expert Guide](#). CESSDA archives (and many more social science data repositories) use the [DDI metadata scheme](#) (DDI Alliance, n.d.). As a result in all CESSDA archives' data catalogues, metadata are structured in the same way.

II. Register yourself as a user

As a registered user you will be able to use more services and functions. Also, registration allows the repository to inform you, e.g. to send you alerts about datasets that you have previously downloaded.

III. Learn how the data repository advanced search functions work

It always pays off to explore the (advanced) search options which the data source of choice offers, e.g:

» **Does it offer truncation?**

Truncation is a searching technique used in databases in which a word ending is replaced by a symbol. Frequently used truncation symbols include the asterisk (*) and a question mark (?).

» **Does it offer wildcards?**

Wildcard symbols can be typed in place of a letter or letters within a keyword if you are not sure of the spelling or if there are different forms of the root word. E.g. The wildcard symbol that should be used is usually an asterisk (*) or question mark (?).

» **Can you use boolean operators?**

Boolean search allows you to combine keywords with operators such as AND, NOT and OR. Be aware that there might be a default boolean operator which is applied standard.

» **Can you use proximity operators?**

A proximity operator is a character or a word (such as NEAR) used to narrow down results by limiting results to keywords which occur within a specific number of words in the content.

» **Can you search in the data themselves or in the metadata only?**

When you can search in the data themselves your query will probably be more exact than when you are searching in the metadata (descriptions of the data) only.

» **Is a controlled vocabulary or thesaurus (predefined keywords to choose from) available?**

A controlled vocabulary or thesaurus is a preselected list of terms used for the description of information, in our case for the description of datasets. The controlled vocabulary used in social sciences is ELSST - European Language Social Science Thesaurus (UK Data Service, 2018).

» **Can you filter results?**

Can you use refinements such as data format, types of analysis, and data availability?

IV. Ask for help

In case you cannot find data (that you know should exist), ask the data service provider. Such user enquiries can help providers develop their collections; if they don't have the data, they may try and find out how to acquire it for future data sharing.

Adjusting your search strategy

When searching for data, you can retrieve too many (mostly irrelevant), too few or no results. How to adjust your search strategy when you do not find what you are looking for?

Assuming that data can be found in the data source of your choice, you can try to rephrase your search query. Some tips:

I. Use appropriate words in appropriate fields

For example:

You are looking for data that reflect the concept of 'postmaterialism'. If you type 'postmaterialism' into the 'question text' field search, the search will retrieve nothing because the word 'postmaterialism' is not included in the wording of the respective questionnaire item. Items dealing with 'postmaterialism' usually ask respondents about particular attitudes (e.g. 'Maintaining order in the nation' or 'Giving the people more say in important political decisions'). You can look for 'postmaterialism' in the 'keywords field' or 'concept field' or try to find other keywords that will be useful in searching, e.g. by consulting the controlled vocabulary or thesaurus, when available.

II. Broaden your scope

Too few or no results? Consider broadening your scope by:

- » thinking about all the terms that relate to your research domain;
- » using fewer search terms in the search field;
- » being less restrictive when using search operators;
- » being less restrictive with using filters.

III. Narrow your scope

Too many and mostly irrelevant results? Consider narrowing your scope by:

- » choosing more detailed search terms;
- » using more words in the search field;
- » being more restrictive when using search operators;
- » being more restrictive by using filters.

Case: Example of using ELSST

A little about ELSST - European Language Social Science Thesaurus

[ELSST](#) (UK Data Service, 2018) is social science multilingual thesaurus that was developed to aid cross-language information retrieval of social science datasets. It contains thousands of terms corresponding to social science concepts, enables users to find terms related to concepts and provides their detailed specification. The thesaurus covers the core social science disciplines: politics, sociology, economics, education, law, crime, demography, health, employment, information and communication technology and, increasingly, environmental science.

ELSST is available in 14 languages: Czech, Danish, Dutch, English, Finnish, French, German, Greek, Lithuanian, Norwegian, Romanian, Slovenian, Spanish, and Swedish. In near future, it will be used for data discovery within CESSDA and will thus facilitate access to data resources across Europe.

Using ELSST

Imagine the situation when you are looking for data that relate to work. You want to find the best search term to effectively find relevant data. You start using ELSST by typing 'work' into the [ELSST search engine](#) (UK Data Service, n.d.a) and find out that:

- 1. The preferred term for 'work' is 'employment'. This means that if you are looking for data relating to 'work', you should use 'employment' as a search term and not 'work'.*
- 2. ELSST shows you a list of 13 language equivalents for 'employment'. You can use the translation of the term when you search in catalogues in other languages. It can also help you when you are simply looking for equivalent terms in other languages.*
- 3. ELSST shows how employment relates to:*

Broader terms (BT)

Concepts that are at an overarching level to 'employment', in this case 'Labour and employment'.

Narrower terms (NT)

More detailed phenomena related to 'employment' in general. In this case 'youth employment', 'job creation' etc.

Related terms

For 'employment', examples of related terms are 'employment policy' and 'right to work'.

4. Select data candidates

When you find data you should ask yourself whether the data seem relevant for your research question. To fully evaluate the suitability/usefulness of data you usually need to scrutinise the documentation described in step 5.

If the data do not seem relevant, ask yourself why you found data that are off-topic? What does this tell you about how you have used the search functionalities of the data source? Return to step 3 or even to earlier steps in the data discovery cycle if necessary and adjust your search strategy.

Expert tips



Check appropriateness of concepts

Bear in mind that the concepts you find in the data should be the same as the concepts from your research question.

Use appropriate indicators

Evaluate how well indicators/questionnaire items/variables apply to your concepts. If you use indicators/variables that do not fit to your concept, your study will lack validity.

5. Evaluate data quality

What quality should you demand from the dataset you have selected as a potential candidate for your research? To determine data quality, familiarise yourself with its content and get a detailed notion about what is in it and what isn't. Think about how the data were collected and ask yourself questions such as:

- » What information was collected, from whom, when and where?
- » Who collected the data and when?
- » Why was the data created? E.g., different purposes for data collection are research, social policy, marketing etc.
- » How was the data collected? You need detailed information about the methodology.
- » How was the data processed? Were there any changes in data? Who adjusted data in what way after it was collected? To which manipulations was the data exposed?
- » Were consistency and logic checks employed? Is the data "clean", i.e. were nonlogical and erroneous values deleted?
- » What quality assurance procedures were used? Did researchers use verified measurement tools?

The information about the data you always need to know is twofold:

I. Project-level documentation

Project-level documentation explains the aims of the study, what the research questions/hypotheses are, the methodologies, instruments and measures being used, etc. Survey data project level documentation also contains detailed information about data collection, respondents, measurement tools, data manipulations etc. It includes user guides, survey questionnaires, interview schedules, fieldwork notes etc.

II. Data-level documentation

You can find more detailed information about what quality you should demand from your data and the accompanying documentation in the [Documentation and metadata](#) paragraph of Chapter 2.

Expert tips



Determine data quality before download

In domain data archives, documentation and metadata is usually publicly available without the need to register as a user. Therefore, you can read it before you download the dataset.

Look out for high quality data documentation

In trusted repositories (such as CESSDA archives), data is accompanied by project-level and data-level documentation. Such documentation can usually be found in documents called user guide, fieldwork notes, technical report or readme files.

Look out for clean data (or clean them yourself)

The data you download from the data repository may be 'clean', i.e. the values of variables have been checked for logical consistency and illogical values have been filtered off. But in some cases, datasets will not be cleaned. In this case you will have to make necessary consistency checks.

Prevent filter bubbles

Don't be satisfied with your preferred method of data discovery. In order to prevent operating in filter bubbles, you should invest enough time in using a mixture of strategies and in visiting multiple sources to find data. In this way you have a chance to locate data which are hard to find or to find data which do not belong to 'the usual suspects' (Gregory et.al., 2018a).

7.2 Data repositories as data resources

In chapter 6, we presented [several types of data publishing routes](#) and types of data repositories. Similarly, when you want to discover research data you will find that they are hosted at different types of data repositories.

Data resources for researching wellbeing: a case study

Bram Vanhoutte is a Research Fellow in Sociology at the Cathie Marsh Institute for Social Research (CMI), The University of Manchester. Bram was appointed as a UK Data Service Data Impact Fellow for 2016-2018 and his research focuses on wellbeing in later life. In the Questions and Answers below, Bram introduces his research and the data he has discovered and uses for his research.

What aspects of wellbeing do you work on?

My research has concentrated on later life wellbeing: how to measure it, how it evolves over time and also how our social trajectories through life influence it. I have just started working on a new research project, [The road to resilience: A comparative life course study](#) (Manchester Institute for Collaborative Research on Ageing, n.d.) which examines the different ways in which people live through adverse events that define ageing in the public eye, such as loss of health, loss of partner and loss of wealth.

What data resources do you use?

For this project, I wanted to use longitudinal data to study how individuals change over time as well as comparative data to examine how different countries compare. Luckily, there are well-established panel studies focused on health and ageing such as the [English Longitudinal Study of Ageing \(ELSA\)](#) (UK Data Service, n.d.b) and its sister studies, the [US Health and Retirement Study \(HRS\)](#) (Health and Retirement Study Survey Research Center, n.d.) and the [Survey of Health and Retirement in Europe \(SHARE-ERIC\)](#) (n.d.). These panel studies allow studying how individuals change over time, how people differ from each other as well as how different countries compare, since they are conceived with international comparisons in mind.

Another enormously useful resource has been [Gateway to Global Aging Data](#) (National Institute on Ageing, n.d.) which provides tools for both searching for relevant search questions but also tools for creating harmonised datasets based on the different studies

Important social science data archives

Social science data archives belong to the category of (trusted) domain repositories. They are important resources for discovering social science datasets (Gregory, et al., 2018a). It is the mission of such repositories to embed data into the research lifecycle in such a way that data are published, shared, discovered and reused. Trusted domain repositories, such as the CESSDA Archives, design their data infrastructures to follow the FAIR (Findable, Accessible, Interoperable and Reusable) data principles (see [chapter 1](#)). Moreover, they:

- » archive and preserve data;
- » offer and manage (mostly online) access to the data;
- » provide complex services focused on data reuse for research, teaching and learning;
- » check data quality and compliance;
- » improve data interoperability, e.g. by accompanying data with rich standardised metadata;
- » maintain (mostly online) data catalogues;
- » seek to add new data to their collections;
- » develop training for data producers and data users.

Important (trusted) domain repositories are:

CESSDA Archives

The [Consortium of European Social Science Data Archives](#) (CESSDA.n.d.b) serves as a platform for development of European integrated data services in social sciences based on wide collaboration among national data archives across Europe. It strives to be a 'one stop shop' for European data.

The CESSDA member archives (CESSDA, n.d.b) provide access to diverse collections of data. Most of the national social science archives dispose of data representing the respective national population. Some large CESSDA archives (e.g. [GESIS](#), n.d.b or [UK Data Service](#), n.d.c) also provide datasets from various international research projects such as [ISSP](#) (n.d). The majority of data provided by CESSDA archives are survey data, i.e. quantitative data, although some archives dispose of qualitative data, e.g., interview transcripts and field notes. The collections of CESSDA archives include data from contemporary research projects as well as older datasets, including longitudinal studies that have been collecting data over decades.

The CESSDA member archives satisfy strict requirements regarding data quality and trustworthiness of the data archive's services and they conform to international standards of data documentation and accessibility. Data services are complemented by different CESSDA products such as services and training activities targeted at data users (researchers), data archives and data professionals.

Search for CESSDA data

The [CESSDA Data Catalogue](#) (CESSDA, n.d.a) is a platform for researchers, where you can search for data from most of the CESSDA archives. Data are not directly downloadable. Instead you will be redirected to the relevant archives for access.



Expert tip: curated directory of international surveys

If you are interested in research data for international comparison, have a look at the [curated directory of international surveys](#) in the online version of this guide.

Out-of-CESSDA European social sciences data archives

CESSDA is continuously widening with the objective to reach a pan-European coverage. However, there are data service providers which are currently not affiliated with CESSDA. A few examples are:

Estonia (ESSDA)

The [Estonian Social Science Data Archive](#) (ESSDA, Eesti Sotsiaalteaduslik Andmearhiiv, n.d.) contains Estonian social science data and survey data, as well as university publications and Estonian radio archival materials. Information is currently only available in Estonian.

Ireland (ISSDA)

The [Irish Social Science Data Archive](#) (ISSDA, n.d.) is Ireland's leading center for quantitative data acquisition, preservation, and dissemination.

Italy (Bicocca Data Archive)

The [Interdepartmental Centre UniData – Bicocca Data Archive](#) (University of Milan-Bicocca, n.d.) is a joint project coming from eight departments of the University of Milano-Bicocca. The project aims to create a center of excellence in data sharing, enhance the secondary analysis of data and promote responsible data use in social, economic and environmental studies.

Lithuania (LiDA)

The [Lithuanian Data Archive for Humanities and Social Sciences](#) (LiDA, Kaunas University of Technology, n.d.) is a virtual centre of expertise in data acquisition, long-term preservation and dissemination established at the Kaunas University of Technology. The archive is promoting access to the national and international collections of digital data in the social sciences and humanities in Lithuania.

Luxembourg (LISER)

[LISER](#) (Luxembourg Institute of Socio-Economic Research, formerly CEPS/INSTEAD) is a Luxembourgish public research institute under the jurisdiction of the Ministry of Higher Education and Research. Its research focus lies in the field of social and economic policy including the spatial dimension. You can visit the LISER data catalogue (LISER, n.d.) to find data.

Poland (ADS)

The [Polish Social Data Archive](#) (ADS, Institute for Social Studies of the University Of Warsaw and Institute of Philosophy and Sociology of the Polish Academy of Sciences, n.d.) is well developed regarding internal standards of data acquisition, archiving and publishing.

Romania (RODA)

The [RODA archive](#) (Romanian Social Data Archive, n.d.) contains data collections accessible for the academic community and the interested public, for secondary and comparative analysis.

Russia (JESDA)

The [Joint Economic and Social Data Archive](#) (JESDA, Higher School of Economics, n.d.) provides free and open access to the results of empirical research in social sciences.

Selected non-European data archives

Here we list some of the well developed data archives in non-European countries to show diversity of data services worldwide:

Canada (Odesi)

[Odesi](#) (Ontario Data Documentation, Extraction Service and Infrastructure, n.d.) is a digital repository for social science data, including polling data. It is a web-based data exploration, extraction and analysis tool that uses the Data Documentation Initiative (DDI Alliance, n.d.) social science data standard

Brasil (CESOP)

[CESOP](#) (Center for Studies on Public Opinion, n.d.) is a center for interdisciplinary research established at the State University of Campinas in 1992. Its central objective is the development of scientific research in the field of political and social behavior.

Israel (ISDC)

[ISDC](#) (Israel Social Sciences Data Center, n.d.) collects, processes, distributes and stores data from different areas in the social sciences. Since its establishment in the late 1970s, the database has developed into a national center.

Japan (SSJDA)

SSJDA (Center for Social Research and Data Archives, Institute of Social Science, The University of Tokyo, n.d.) is a comprehensive archive of social science data concerning Japan.

South Korea (KSDC)

[KSCD](#) (Korean Social Science Data Center, n.d.) was established in November, 1997 to build a new system of managing comprehensive sources of social science data.

Taiwan (SRDA)

[SRDA](#) (Survey Research Data Archive, n.d.) was founded in November 1994 by the Center of Survey Research (CSR), formerly the Office of Survey Research. SRDA engages in the systematic acquisition, organisation, preservation, and dissemination of academic survey data in Taiwan.

Australia (ADA)

[ADA](#) (Australian Data Archive, n.d.) provides a national service for the collection and preservation of digital research data.

US (ICPSR)

[ICPSR](#) (Inter-university Consortium for Political and Social Research, n.d.) in the United States, has many datasets on American society, but its scope is worldwide, as its member institutions come from all parts of the world.

Other important data repositories

Here we list some important institutional or project data repositories:

EUROSTAT

A very important source of data on European and EU countries is EUROSTAT, the statistical office of the European Union. EUROSTAT key task is to provide statistics at European level that enable comparisons between countries and regions. It provides access to data in two categories:

» **The Eurostat data collection**

The Eurostat data collection (in aggregate form) can be accessed [here](#) (European Commission, n.d.b).

» **Eurostat microdata**

Eurostat microdata (including the [European Union Statistics on income and living conditions](#) (European Commission, n.d.a) can be accessed under specific conditions, frequently through some form of secure access (especially in case of confidential data). More information on how to access Eurostat microdata can be found in the publication '[How to use microdata properly](#)' (European Commission, 2018).

European Union Open Data Portal

[European Union Open Data Portal](#) (EU ODP, European Commission, n.d.c.) gives access to open data published by EU institutions and bodies.

OECD statistical data

[OECD iLibrary](#) (OECD, n.d.) is the online library of the Organisation for Economic Cooperation and Development featuring its books, papers and statistics and is the gateway to OECD's analysis and data.

United Nations (UNdata)

[UNdata](#) (United Nations, n.d.) is a web-based data service for the global user community. It brings international statistical databases within easy reach of users through a single-entry point. Users can

search and download a variety of statistical resources compiled by the United Nations statistical system and other international agencies.

UNESCO

UNESCO Institute for Statistics offers [data for the Sustainable Development Goals](#) (UNESCO, n.d.).

UNICEF data

[UNICEF data](#) (UNICEF, n.d.) monitors the situation of children and women worldwide.

World Bank Open Data

[World Bank Open Data](#) (The World Bank, n.d.) offers free and open access to global development data.

European longitudinal research projects

- » [The European Social Survey](#) (ESS-ERIC, n.d.);
- » [Generations & Gender Programme](#) (GGP, n.d.).

European diversity

Data archives for social sciences differ considerably between European countries. Below you find an example of a 'small' archive (CSDA) and a 'large' archive (UKDS).

CSDA

The [Czech Social Science Data Archive](#) (CSDA, n.d.) was founded in 1998 as a department of the Institute of Sociology in Prague.

The large majority of the CSDA collection consists of data from sociological surveys. The data collection is gradually growing (in 2018, over 800 data sets are available) and expanding beyond the frontiers of sociology. With a few exceptions, research data cover only the area of Czechia. Only a small part (less than 10 %) of the data collection is in English.

UKDS

[UK Data Service](#) (n.d.d) provides access to the UK's largest collection of social, economic and population data. UKDS also supports users with training and guidance.

The data collection includes major UK and cross-national surveys, including many government sponsored surveys and longitudinal studies and several cohort studies following individuals born in 1958, 1970 and 2000. There is data from the UK Census from 1971 to 2011 and qualitative data collections containing in-depth interview transcripts, diaries, anthropological field notes, etc.

The UK Data Service has an online repository called Reshare (UK Data Service, n.d.e) for researchers to archive, publish and share research data. Reshare is an important tool in helping researchers to comply with the data archiving requirements from the UK's Economic and Social Research Council.

Expert selections of data resources

In the online version of this guide, CESSDA-experts highlight key data resources for several research topics and show you how to access the data. Maybe you can find something for your research interests.

- » [Data resources for ageing](#)
Key European data resources for research related to ageing and its effects on individuals and society.
- » [International comparisons](#)
Interested in research data for international comparison? Have a look at our directory of international surveys.
- » [Other curated data sources](#)
CESSDA prepares data discovery materials, selections of data resources and organises data discovery events.

7.3 Resources for social media data

Social media data come from various resources, such as Facebook, Twitter, Reddit, Instagram or YouTube. The elements of social media data may be:

- » individual tweets, comments on Facebook, Twitter or Reddit etc.,
- » visual content, such as photos or videos,
- » network connections between network users (friend connections, groups),
- » data on ratings and/or interests (preferences or likes).

Social media data are available to researchers, but their availability is restricted by companies that own respective social media platforms (Facebook, Twitter, etc.). Restricted availability of social media data represents serious obstacle for more intensive application of social media data in social research.

There are several reasons for the limited availability of social media data. One of them is legal and deals with the social media content's copyright. The users have copyright for their own content (e. g. Tweets or Facebook posts) and by signing terms of use they give the social media platform a license to use the content for various purposes. The use of the social media data for third parties (private companies, academic researchers etc.) is restricted in the terms of use. This constrains the researchers (and data archives) in using, storing and sharing the data. A good source of guidance on social media data preservation both for researchers and repositories is Thomson, S.D. (2016) "[Preserving Social Media](#)".

One of other reasons for the limited availability of social media data lies in the ethics. Researchers and data archivist must care about the protection of personal information of the social media users.

Platforms as social media data sources

Social media data can be obtained through the application programming interfaces (APIs) of the social media platforms. However, these APIs usually restrict the type and amount of data you can collect. If researchers request large amounts of data through APIs, they might not get the complete data but samples. Often it is not fully transparent how these data are sampled.

For those who are not able to handle APIs for downloading the data, there are commercial subjects that sell social media data, such as [Gnip](#) (acquired by Twitter Inc. in 2014) or [DataSift](#), but these usually have high costs.

Social Media Data in European Data Archives

According to the results of a survey carried out among European social science data archives for the [SERISS project](#) in June 2019, only two CESSDA archives store and disseminate social media data so far: GESIS and UK Data Service (UKDS) offer their users limited collection of social media data, Facebook data, geo-coded Twitter data, and specific subsets of Wikipedia. In particular, UKDS holds several Twitter data sets (20 collections of Twitter communication (tweets' IDs, timestamp, hashtags).

Currently, several CESSDA archives plan strategies to overcome legal and technical issues related to social media data archiving and sharing as they see it as important area.

General repositories

[Zenodo](#), [Harvard Dataverse](#) or [Fig share](#) hold limited but increasing number of social media datasets. These repositories obtain data through self-archiving i.e. without archive taking care over data and metadata quality.

Field specific and thematic social media data sources

There exist several projects and institutions that ingest and store social media data on various topics. Some of them are:

- » The [CrisisLex](#) is a repository of crisis-related social media data and tools.
- » The Schlesinger Library on the History of Women in America has created dataset "[#metoo Digital Media Collection](#)".
- » [Stanford Network Analysis Project](#) is a repository for data from internet-based social networks.
- » [TweetsKB](#): A Public and Large-Scale RDF Corpus of Annotated Tweets
- » The [Documenting the Now](#) catalog of Twitter data collections
- » The [TweetSets](#) collection of Twitter data sets

7.4 Access, use and cite data

Once you find suitable data for your purpose and you've checked data quality (See the paragraph [The process of data discovery](#)), how can you get it? The steps below emphasize a number of aspects that you may encounter along the way:

1. Check the terms and conditions of access and use

Different access arrangements may be in place for any data collection, especially those containing more detailed, sensitive or confidential data. Generally, the following access options exist:

» **Open Data**

Anyone can freely access, use, modify, and share for any purpose.

» **Open after registration**

The user must register and provide the required information. The information often includes personal data, institutional affiliation, and purpose of use. No specific conditions of access are required.

» **Open under specific terms and conditions**

For example:

- » Access to a data collection requires permission from the data depositors or data owners;
- » 'Scientific use files' are available only for academic research and education;
- » Data use is limited to 'non-commercial use only';
- » Sensitive and confidential data are available only under strict conditions of use and security measures.

» **Access to metadata only**

Many data files are inaccessible for different reasons. Even if data files are inaccessible, relevant metadata may still be available in repositories and information obtained from it may be also helpful for your research.

» **Embargo**

Some repositories contain data under embargo, i.e. after a specified time period (e.g., 6 or 12 months) the data is released for public use.

Also see Chapter 6 for information on [access categories from the viewpoint of the data depositor](#).

Examples

Different categories of access at GESIS data archive

The access and use conditions for different types of data may vary, even in the same repository. For example, the following different access categories are provided to data by [GESIS](#) (GESIS. n.d.b), Germany:

» **Category 0**

Data and documents are released for everybody.

» **Category A**

Data and documents are released for academic research and teaching.

» **Category B**

Data and documents are released for academic research and teaching, if the results are not published. If any publication or further processing of the results is planned, permission must be obtained by the Data Archive.

» **Category C**

Data and documents are only released for academic research and teaching after the data depositor's written authorization. For this purpose the Data Archive obtains written permission with specification of the user and the analysis intention.

Terms and conditions of access at UK Data Service

UK Data Service [uses the following access categories](#) (UK Data Service, n.d.g) to support access to its large collection of data from various sources, including the UK Office for National Statistics:

» **Standard access**

Applies to the majority of UKDS data and only requires user and project registration. These data are fully anonymised.

» **Special Conditions**

Are usually specified by the data owners and users agreement on them is required during the download/ordering process.

» **Special Licence**

Used for data collections containing more detailed (and therefore potentially disclosive) data such as smaller scale geographical information. If you apply for specially licence data, you need to provide more detailed information about the intended use of the data using a set of Special Licence forms.

» **Secure Lab (Controlled data)**

Provides secure access to data that are too detailed, sensitive or confidential to be made available under other arrangements such as a Special Licence. To use the Secure Lab, you need to complete a special application and attend a training course. Data accessed in this way cannot be downloaded. Once researchers and their projects are approved, they can analyse the data remotely or by using the UKDS Safe Room.

Scientific use files and public use files at Eurostat

In order to protect the anonymity of individual persons or businesses, access to confidential microdata at Eurostat is restricted. Most of Eurostat's microdata is accessible only in the form of so called scientific use files (SUFs) for scientific purposes only.

The access is based on [a complicated system of accreditation](#) (European Commission, n.d.d). In addition, there are also public use files (PUFs) or public microdata which are made available to public. These files are prepared in such a way that individual entities cannot be identified. However, this de-identification is accompanied by a loss of informative value in the data.

Licence agreements

Data files may be copyrighted work and therefore subject to copyright specified in the terms and conditions of use. Nowadays, the agreement with conditions of use is usually available on-line, but a written agreement may be required at some repositories or under some circumstances. Sometimes, especially for datasets classified as Open Data, [CC licences](#) (Creative Commons, 2017) are used to facilitate access to data. For more information on licence agreements read the [Licencing your data](#) paragraph of Chapter 6.

2. Consider possible ways of access and use of the data

Nowadays, data organisations and projects are increasingly offering tools for on-line data analysis in addition to direct downloads. Sometimes different ways of access are offered to the same data file by different repositories.

Examples of ways of access are:

» **Direct download**

Direct download is the easiest way to get the data. However, you should consider the availability of appropriate analytical software, the structure and formats of the dataset. Experienced analysts usually prefer direct downloads as capabilities of on-line tools are often limited to very basic analytical methods.

» **Online analysis**

The advantages of online analysis is that you do not need your own specialised software. In addition, especially if the dataset has a complicated structure, the online tool may be a source of higher operability. It may allow easier orientation in a complex database, selecting, linking or merging of its different sections, selecting correct weighting factors, etc.

Tools for on-line analysis are available at, e.g.:

- » [The European Social Survey](#) (ESS ERIC (n.d.b));
- » [World Values Survey](#) (WVS, Institute for Comparative Survey Research (n.d.b));
- » [IEA IDB Analyzer](#) (IEA. n.d.).

» **On site access in the safe room**

Secure data centres with safe rooms provide access to highly sensitive and confidential data under strict security measures. Researchers are required to apply for accreditation, travel to the location of the centre, and work with the data in the safe room. For example, see the [description of the Safe Room at the UK Data Service Secure Lab](#) (UK Data Service, n.d.f).

» **Secure remote-execution system**

A secure remote-execution system is an alternative way to make confidential data accessible. The data user has access to rich metadata, but not directly to the dataset. Instead, statistical programs of intended analysis are submitted and on return aggregated results are obtained. For example, [LISSY](#) (LIS, n.d.) allows researchers to access microdata from the [Luxembourg Income Study](#) (LIS, n.d.b) and the [Luxembourg Wealth Study](#) (LIS, n.d.c). Users submit their statistical programmes written in R, SAS, SPSS or Stata via the Job Submission Interface or via email. LISSY automatically processes the jobs and returns back aggregated results within few minutes.

Case: Using NESSTAR for data discovery

As we have noted, there are differences in ways of data presentation and functionalities of search among individual repositories. Some CESSDA archives use [NESSTAR](#) (NSD, n.d.) software. NESSTAR is a software system for publishing and presenting data on the Web. Some data services use NESSTAR as their main tool for searching and accessing data while others have a main catalogue and provide NESSTAR as an additional tool. NESSTAR enables online data browsing and analysis. You can also download tables, graphs, data files and study descriptions. NESSTAR help pages, accessible by clicking the question mark at the top of the screen, include helpful guidance. In NESSTAR, you can use advanced search.

3. Consider the costs and the time it takes to access data

Not all available data can be accessed free of charge. Even if the principles of open access to research data are applied, coverage of the marginal costs of access may be required from data users. For specific types of access, the expenses may be considerable. E.g. when you have to cover travel expenses to gain access at secure data centres.

In addition to the costs associated with access, it can also take time to gain access. For example, administration of requests and authorisation procedures for access to confidential and sensitive data is often time-consuming.

4. Consider the format of data and metadata

If you download data, it does not mean they are always available in the format you need. Some tips:

- » Keep in mind that raw research data may have a specific structure and their efficient processing and analysis may require specialised software and skills.
- » Data services often offer downloads in several different formats. Sometimes, however, only one format is available. If it is a current, standard analytical software format, there is usually no problem. In contrast, old proprietary formats can cause significant difficulties. An overview of data formats and more information about format conversion [is available in Chapter 3](#).

This Expert Guide does not focus on data processing for purposes of data analysis. However, the following chapters can help you in understanding your data and their preparation for analysis:

[Chapter 2. Organise & Document](#)

[Chapter 3. Process](#)

Challenges in using data

After downloading the data, you will have to make the data suitable for reuse. The case study below shows that the challenges you may encounter before you can actually start using and analysing the data may be complex.

Case study: Data for a replication study

Kristyna Bašná works at the Institute of Sociology of the Czech Academy of Sciences (n.d.). She needed data for a replication study. How did she discover, access and use such data?

What kind of data were you looking for?

My research focuses on the relations between structural properties of states, civic culture attitudes and change in the level of democracy. My research is a replication of a well-known paper written by Muller and Seligson (1994) who did a cross-national analysis on 27 countries and concluded that civic culture does have an important influence on the level of democracy.

To be able to replicate this analysis I needed data that would allow for cross-national and longitudinal comparison. At the same time the data should be comparable with the data used in the data analysis of Muller and Seligson. Data such as GDP per capita, level of democracy or Gini coefficient, are relatively easily accessible. However, it was much harder to find data with variables identical to the variables which were used by Muller and Seligson to measure civic culture. Yet this was exactly what I needed in order to be comparable.

How did you locate suitable data resources?

I decided to search all the different cross-national public opinion survey databases and look for the exact same question that was used by Muller and Seligson (1994). In the end I was able to find data on 85 countries ranging from 1981 to 2015, in total having 337 country years. I downloaded the data on civic culture from openly accessible resources such as:

*The [European Values Study](#) (GESIS, n.d.c);
[Eurobarometer](#) (GESIS, n.d.d);
[World Values Survey](#) (Institute for Comparative Survey Research, n.d.);
[LAPOP](#) (Latin American Public Opinion Project, n.d.).*

I also downloaded:

*[data on democracy from Polity IV](#) (Center for Systemic Peace, n.d);
 data on GDP per capita and Gini coefficient from [the World Bank](#) (The World Bank, n.d.).*

What challenges did you encounter before you could use the data?

Downloading data from multiple resources is not a straightforward task because most databases use different coding. It is therefore essential to combine the data from multiple sources correctly and with the utmost care, because variables names and country names may differ, data may be missing and different types of weighting may have been used. In my case, I did not need data about individuals, but data collapsed by country and year. That is why for each database I first collapsed the data (using weights) keeping only the variables that I needed for my analysis.

In the second step, I made sure that the country names were identical in each of my data resources. I had to recode a number of countries because some surveys used very different coding. I also had to ensure that the variable on civic culture was identically coded in all of the different data resources, which was fortunately the case. Finally, I have merged all the different datasets into one big data file, which I then used for my quantitative analysis and for the replication of the Muller and Seligson (1994) article.

Citing data

After you have used research data you may want to publish about the work you have done. In this case, you should always cite research data. Research data may be subject to intellectual property rights. However, citing data is usually included in the terms and conditions for the use of data. The obligation to properly acknowledge any research work, including the work invested into development of databases, also logically follows from research ethics.



Expert tip: Use a persistent identifier

In citation always use persistent identifiers (DOI – Digital Object Identifier) if available. It promotes findability and accessibility of data.

Minimal data citation

The [minimal data citation recommended by DataCite](#) (Datacite, n.d.b) is:

Creator (PublicationYear). Title. Publisher. Identifier

DataCite recommends including information about two optional properties, Version and ResourceType (if applicable):

Creator (PublicationYear). Title. Version. Publisher. ResourceType. Identifier

Examples of data citations

Political Party Database, 2011-2014 (APA)

Webb, P., Scarrow, S., Poguntke, T. (2017). Political Party Database, 2011-2014. [data collection]. UK Data Service. SN: 8265, <http://doi.org/10.5255/UKDA-SN-8265-1>

Political Party Database, 2011-2014 (Harvard)

Scarrow, S., Webb, P., Poguntke, T., 2017, Political Party Database, 2011-2014, [data collection], UK Data Service, Accessed 17 October 2018. SN: 8265, <http://doi.org/10.5255/UKDA-SN-8265-1>

European Working Conditions Survey, 2015 (APA)

European Foundation for the Improvement of Living and Working Conditions. (2017). European Working Conditions Survey, 2015. [data collection]. 4th Edition. UK Data Service. SN: 8098, <http://doi.org/10.5255/UKDA-SN-8098-4>

European Working Conditions Survey, 2015 (Harvard)

European Foundation for the Improvement of Living and Working Conditions, (2017). European Working Conditions Survey, 2015. 4th Edition. UK Data Service. [data collection]. <http://doi.org/10.5255/UKDA-SN-8098-4>

Why should I cite data?

Have a look at the video below (ICPSR, 2018) to learn about the benefits:

<https://www.youtube.com/watch?v=jiCZKV-alCQ>

More about data citation can be found in [Chapter 6](#) or, e.g., in the [IASSIST Quick Guide to Data Citation](#) (IASSIST, 2012).

7.5 Adapt your DMP: part 7



This is the seventh and final 'Adapt your DMP' section in this tour guide. After working on this chapter, you should be able to plan for data discovery. To adapt your DMP, consider the following elements and corresponding questions:

Identification of needs

- » Do you plan to use existing data for your research?
- » What is the purpose for which you need the data?
- » What do you want to learn from the data?
- » What type of data do you need?

Search for data

- » Do you know where the data may be located?
- » How do you plan to search for the data?

Evaluation of data quality

- » What is the minimal required quality of the data (in terms of origin, contents, scope, size, methods, etc.)?
- » How do you plan to evaluate data quality (evaluation of metadata, tests, analysis, comparisons)?

Gaining access to data

- » What are the (expected) terms and conditions for data access and use?
- » What is the (expected) process for gaining access to the data?
- » What is the (expected) time-span of the process for gaining access to the data?
- » What are the (expected) costs for data access and use?

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can [view and download the checklist as pdf](#) (CESSDA, 2018a) or [editable form](#) (CESSDA, 2018b), and keep them as a reference while you are studying the contents of this guide.

Sources and further reading

[Please see the online version of this guide.](#)

Contributors

The original version of the Data Management Expert Guide was created for CESSDA ERIC by a number of its service providers' experts at: [ADP](#), [AUSSDA](#), [CSDA](#), [DANS](#), [FORS](#), [FSD](#), [GESIS](#), [NSD](#), [SND](#), [So.Da.Net](#) and [UKDS](#) and is illustrated and edited by [Verbeeldingskr8](#). The authors are mentioned by their names at the end of the relevant chapter(s). DANS led the creation of this expert guide.

In 2018 and 2019 additional content was contributed by different service providers' experts. We list below the complete list of contributors.

List of contributors

Bezjak Sonja, Slovenian Social Science Data Archives (ADP)
Bishop Libby, GESIS Leibniz Institute for Social Sciences
Bradić-Martinović Aleksandra | Data Centre Serbia for Social Sciences (DCS)
Brandby Eira, Swedish National Data Service (SND)
Braukmann Ricarda, Data Archiving and Networked Services (DANS)
Breuer Johannes, GESIS Leibniz Institute for Social Sciences
Buckley Jennifer, UK Data Service
Chylikova Johana, Czech Social Science Data Archive (CSDA)
Fält Katja, Finnish Social Science Data Archive (FSD)
Henriksen Gry, Norwegian Center for Research Data (NSD)
Jakobsson Ulf, Swedish National Data Service (SND)
Krejčí Jindrich, Czech Social Science Data Archive (CSDA)
Kondyli Dimitra, Greek research infrastructure for the social sciences - So.Da.Net
Kudrnacova Michaela, Czech Social Science Data Archive (CSDA)
Leenarts Ellen, Data Archiving and Networked Services (DANS)
Lundgren Malin, Swedish National Data Service (SND)
Oliveau Sébastien | French infrastructure for Social Science and Humanities PROGEDO
Peuch Benjamin, Social Sciences Data Archive (SODA)
Recker Jonas, GESIS Leibniz Institute for Social Sciences
Masten Sergeja, Slovenian Social Science Data Archives (ADP)
Stam Alexandra, Swiss Centre of Expertise in the Social Sciences - FORS
Summers Scott, UK Data Service
Van den Eynden Veerle, UK Data Service
Vávra Martin, Czech Social Science Data Archive (CSDA)
Vipavc Brvar Irena, Slovenian Social Science Data Archives (ADP)
Watteler Oliver, GESIS Leibniz Institute for Social Sciences