

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

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)

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



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**

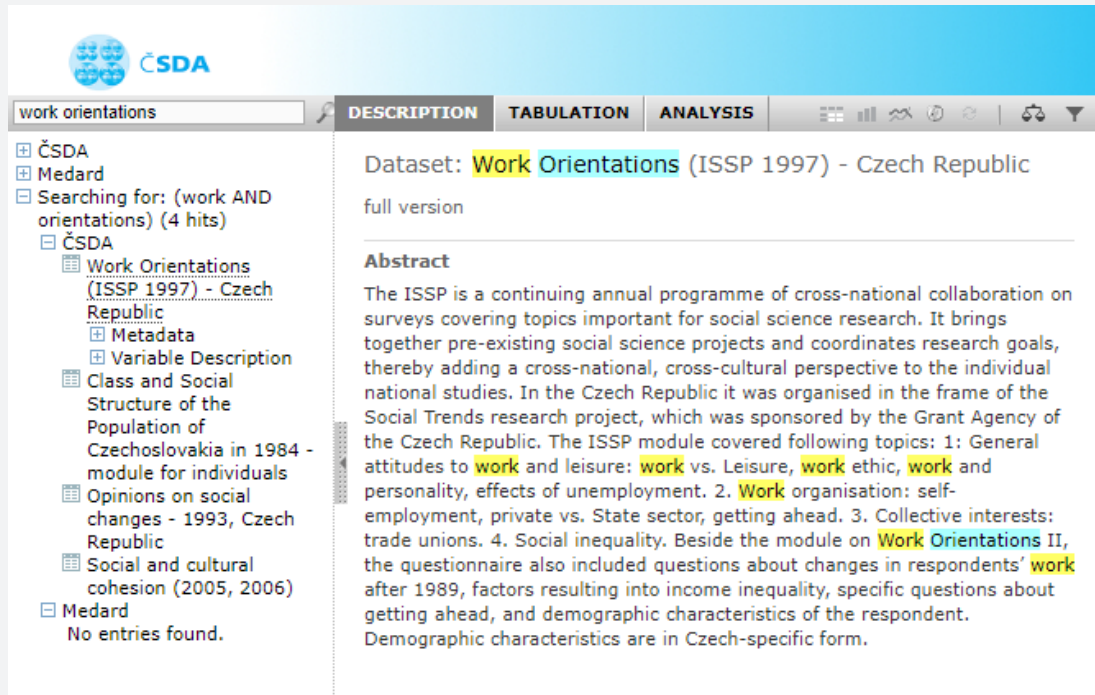
Jana begins with a review of scholarly literature and looks for data used by other researchers. She has read a number of studies on the topic before and now she reviews the 'research method' and 'data' sections in articles to learn about the data and the data resources. The authors mention several publicly 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**

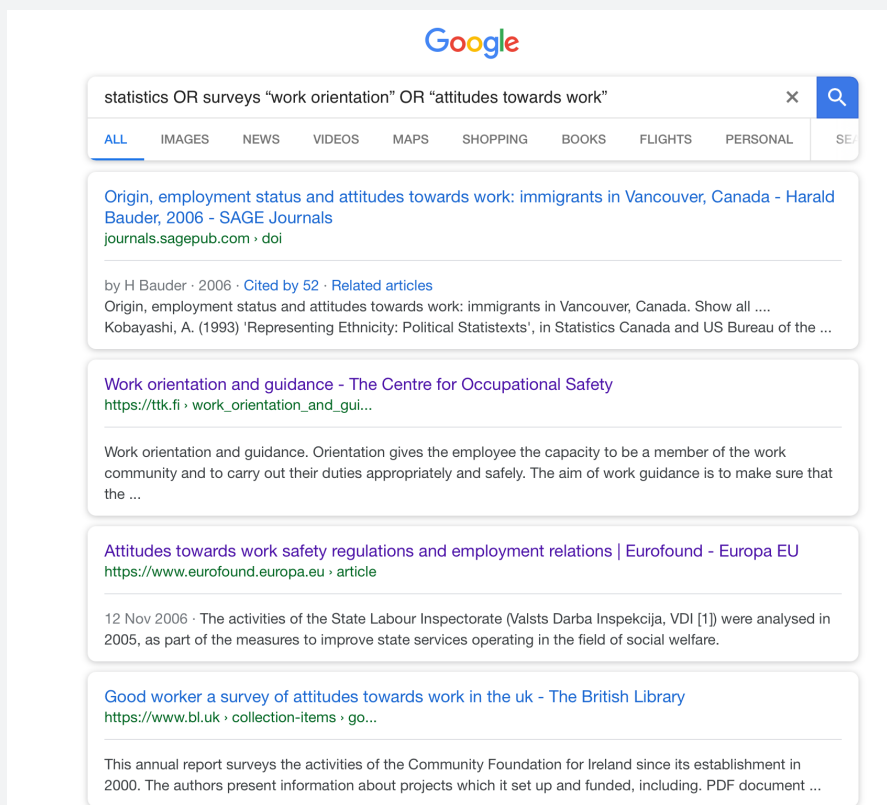
Jana searches the GESIS data catalogue also for other data on work orientations. Besides the ISSP, there are over a hundred other studies. However, they are not comparative by nature or the topic of work orientations is not so central. Jana decides to visit the Czech Social Science Data Archive (CSDA, 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. On the left, a search bar contains 'work orientations'. Below it, a tree view shows the search results: 'ČSDA', 'Medard', and 'Searching for: (work AND orientations) (4 hits)'. Under 'Searching for...', there are four hits: 'ČSDA', 'Work Orientations (ISSP 1997) - Czech Republic', 'Class and Social Structure of the Population of Czechoslovakia in 1984 - module for individuals', and 'Opinions on social changes - 1993, Czech Republic'. The 'Work Orientations (ISSP 1997) - Czech Republic' entry is selected, showing its metadata: 'Metadata', 'Variable Description', and 'Social and cultural cohesion (2005, 2006)'. The main panel displays the dataset description: 'Dataset: Work Orientations (ISSP 1997) - Czech Republic', 'full version', and an 'Abstract'. The abstract describes the ISSP as a continuing annual programme of cross-national collaboration on surveys covering topics important for social science research. It mentions that 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. It also mentions that the questionnaire 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 for the query 'statistics OR surveys "work orientation" OR "attitudes towards work"'. The search bar at the top contains the query. Below the search bar, there are tabs for 'ALL', 'IMAGES', 'NEWS', 'VIDEOS', 'MAPS', 'SHOPPING', 'BOOKS', 'FLIGHTS', 'PERSONAL', and 'SERIES'. The search results are displayed in a list of cards. The first card is titled 'Origin, employment status and attitudes towards work: immigrants in Vancouver, Canada - Harald Bauder, 2006 - SAGE Journals' and includes a link to 'journals.sagepub.com'. The second card is titled 'Work orientation and guidance - The Centre for Occupational Safety' and includes a link to 'https://ttk.fi'. The third card is titled 'Attitudes towards work safety regulations and employment relations | Eurofound - Europa EU' and includes a link to 'https://www.eurofound.europa.eu'. The fourth card is titled 'Good worker a survey of attitudes towards work in the uk - The British Library' and includes a link to 'https://www.bl.uk'. Each card also includes a brief description of the content.

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.

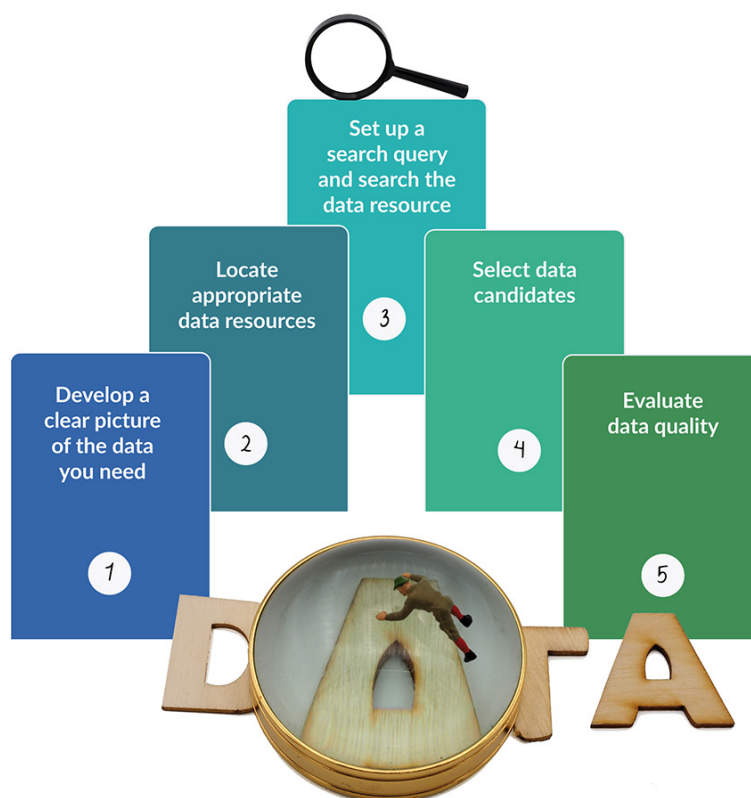
Enhance data quality and foster innovation

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.

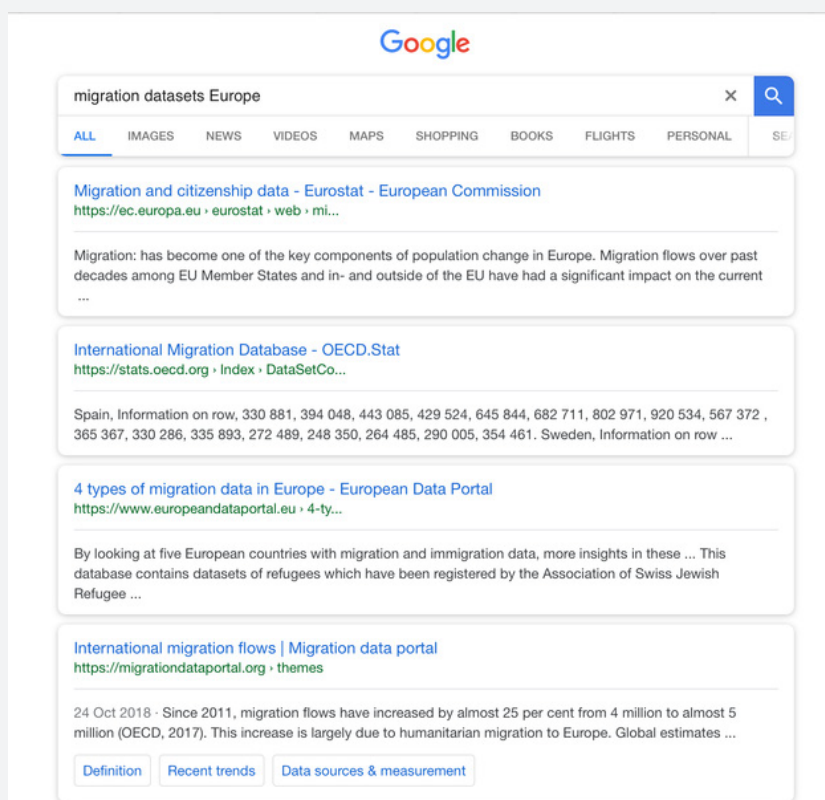
- » OpenAIRE Explorer (OpenAIRE, n.d.) provides a searchable registry of open access compatible repositories.
- » With OpenDOAR (Jisc, n.d.), the Directory of Academic Open Access Repositories, you can browse over 3500 academic open access repositories. It enables users to identify, browse and search 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:

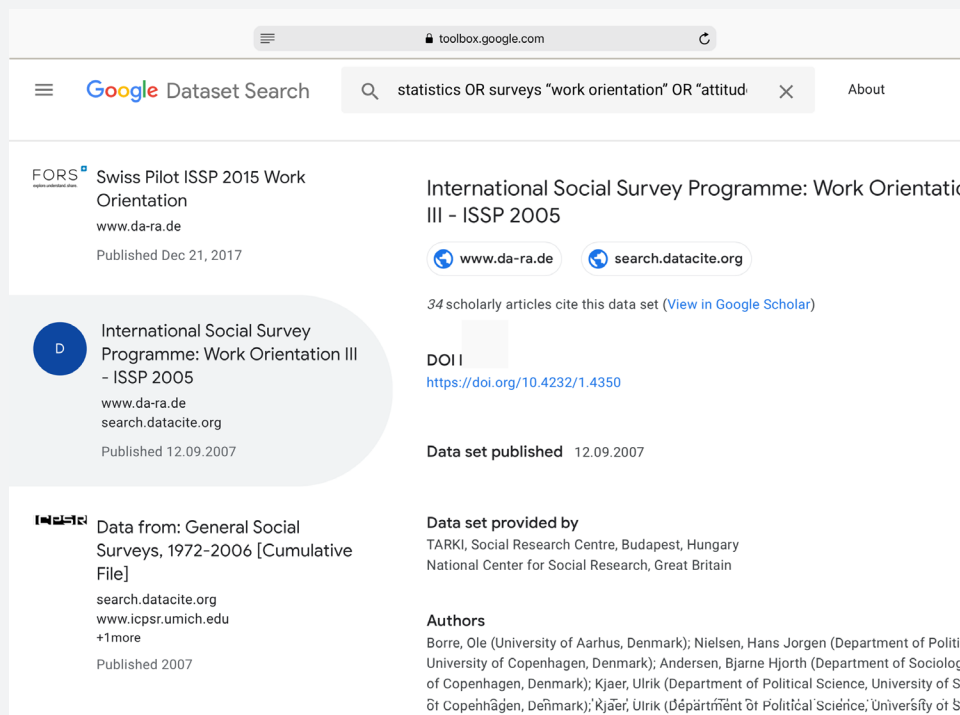
Google

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.



(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 Resource Type (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-aIC0>

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.