Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Welcome

Cristina Magder / UK Data Service

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Workshop netiquette

The presentations will be recorded and shared with the participants shortly after the event.

Please ensure you are muted during the presentations.

Any questions can be added in chat, or anonymously via Padlet.

All slides and materials will be made available on Zenodo for future use under a CC BY licence.

Today’s Speakers and Facilitators

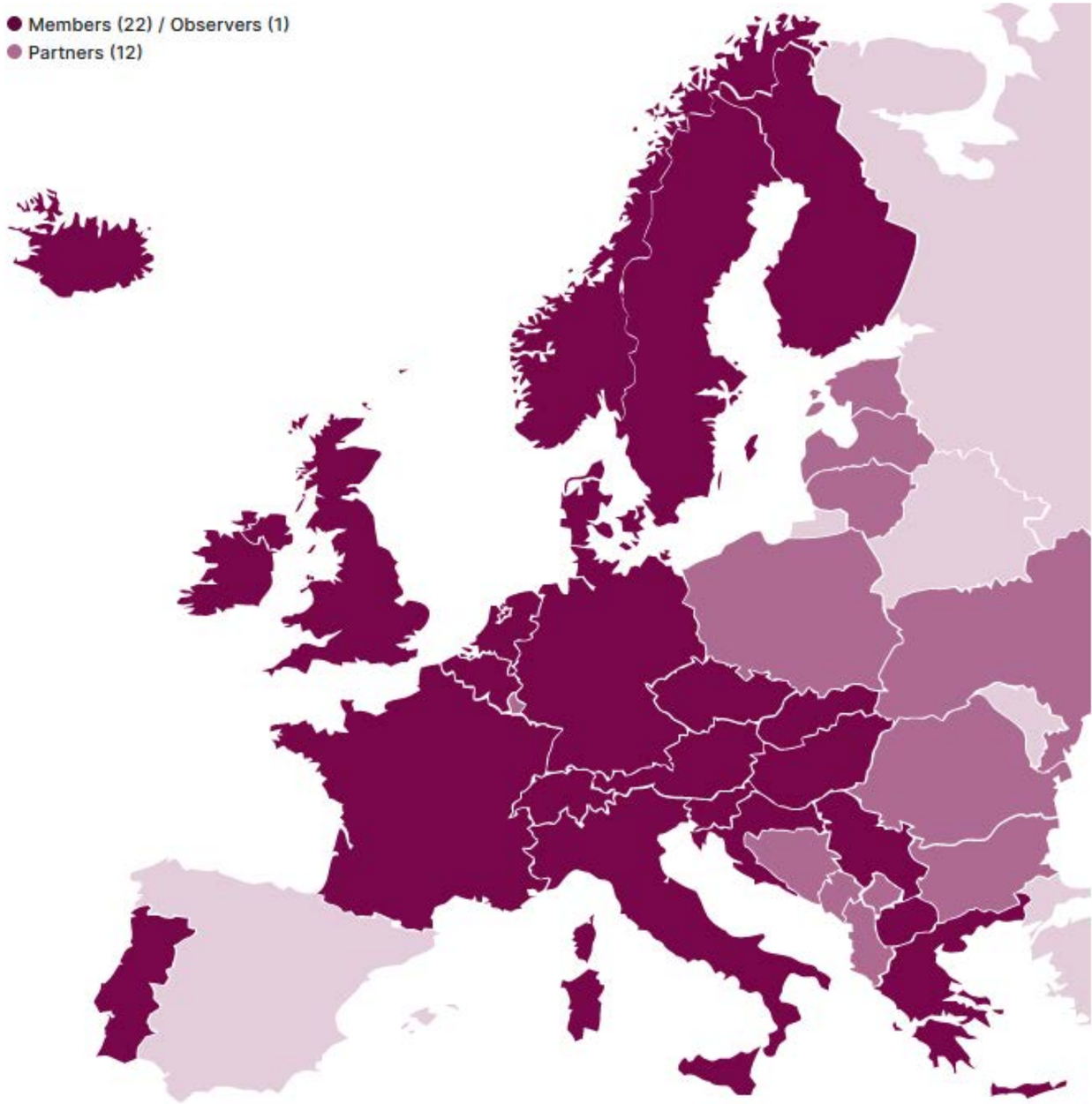
- Cristina Magder – Data Collections Development Manager @UK Data Service, UK Data Archive
- Johana Chylíková – Methodological research, data management @Czech Social Science Data Archive
- Dimitri Prandner – Senior research associate, data acquisition, data management @Austrian Social Science Data Archive
- Marijana Glavica – Acting head @Croatian Social Science Data Archive
- Otto Bodi-Fernandez – Senior research associate @Austrian Social Science Data Archive
- Oliver Watteler – Deputy head of Data Acquisitions and Access @GESIS Leibniz Institute for the Social Sciences
- Maureen Haaker – Senior Qualitative Data and Training Officer @UK Data Service, UK Data Archive
- Vlad Voina – Data Integrity Officer @UK Data Service, UK Data Archive

Our vision is that the provision of **access to social science data** and metadata is vital – for both science and society.

For this we must offer **services to data producers** to easily describe and store their data – if needed in a secured environment.

We will adhere to the **FAIR** (Findable, Accessible, Interoperable, Reusable) data principles to make data findable and provide information about the data, where they are, how they can be accessed.

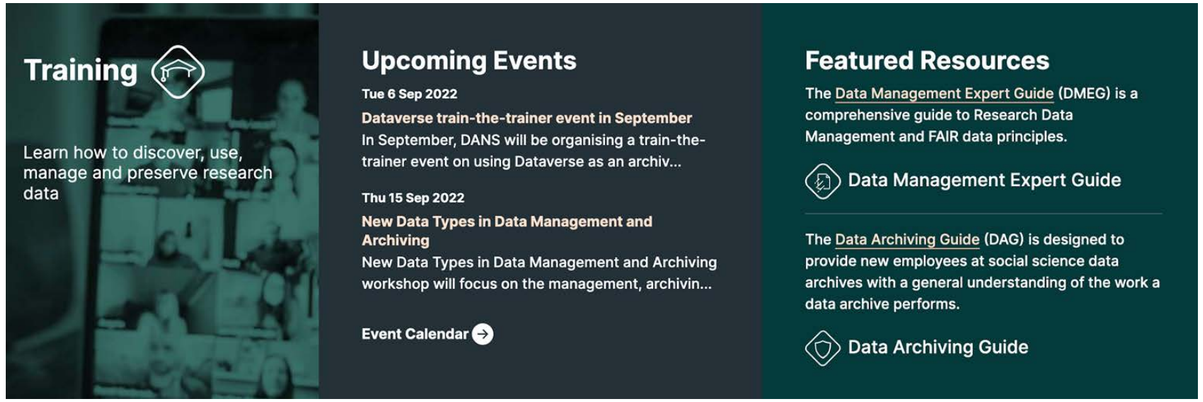
We will also focus on providing **training** and enabling the transfer of expertise and sharing of knowledge on data, as well as relevant rules and regulations.



Mission of CESSDA

- Provide a distributed and sustainable **research infrastructure**
  - **enabling** the research community to conduct high-quality research in the social sciences,
  - contributing to the production of effective solutions to the **major challenges facing society today**.
- Facilitate **teaching and learning** in the social sciences.

Services Supporting Research



Trust and Standards

Each national Service Provider works towards achieving the Trustworthy Digital Repository (TDR) standard selected by CESSDA: the CoreTrustSeal.

The CoreTrustSeal defines repository requirements in terms of organisational infrastructure, digital object management, technology and security. Trust is essential between CESSDA Service Providers and with their data depositors and users.

Since CESSDA selected the internationally recognised CoreTrustSeal it has been acknowledged as important for enabling FAIR (Findable, Accessible, Interoperable, Reusable) data and has become the recommended certification approach for data repositories within the European Open Science Cloud (EOSC).

CESSDA Digital Tools

CESSDA Data Catalogue

Search tens of thousands of social science research studies from our European Service Providers.

ELSST Thesaurus

The European Language Social Science Thesaurus is a broad-based multilingual thesaurus for the social sciences.

European Question Bank

The EQB is a cross-national question bank for social science and humanities research.

Vocabulary Service

Search, browse and download controlled vocabularies in a variety of languages.

Resource Directory

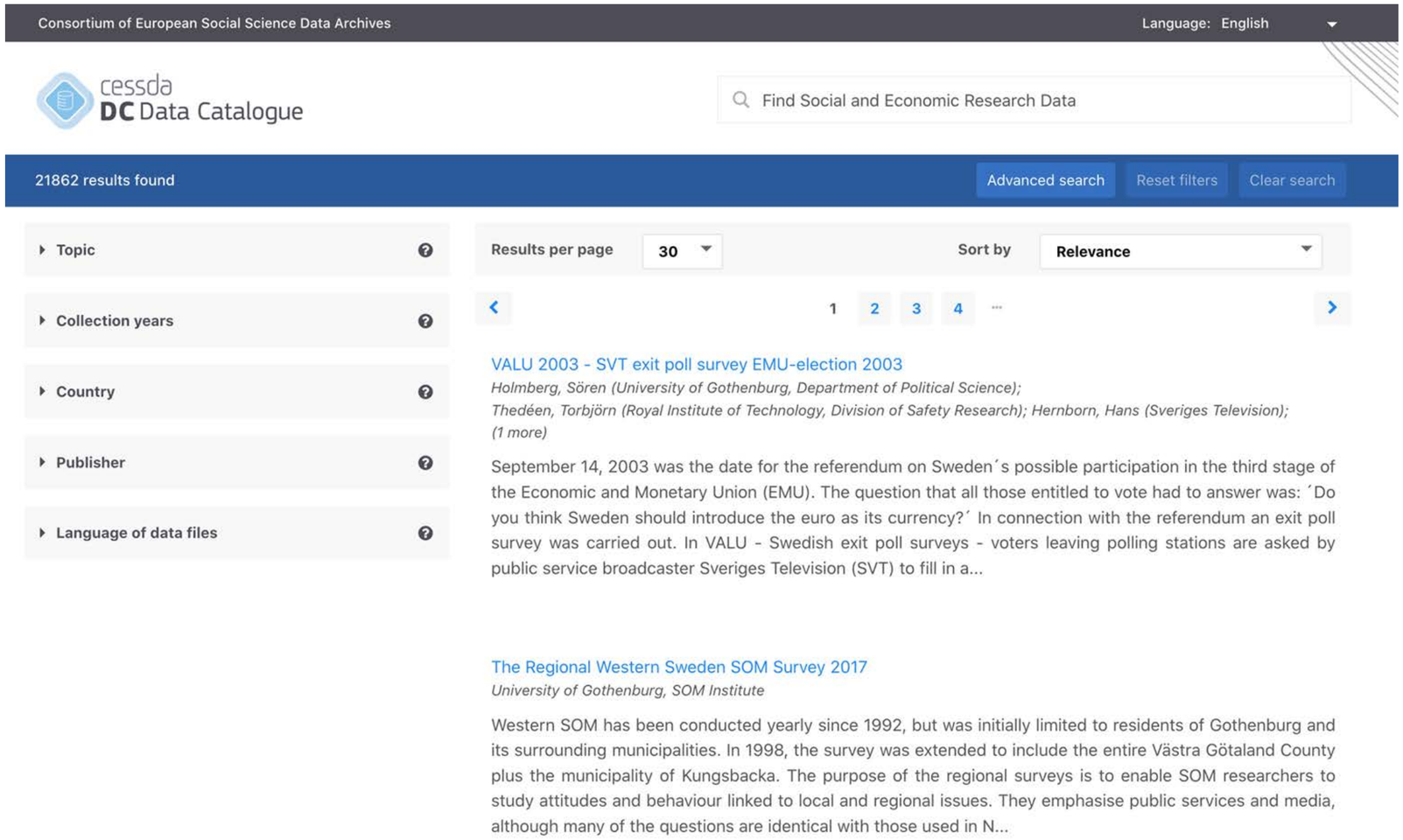
Access resources for data archives and data professionals from CESSDA, its Service Providers and partners.

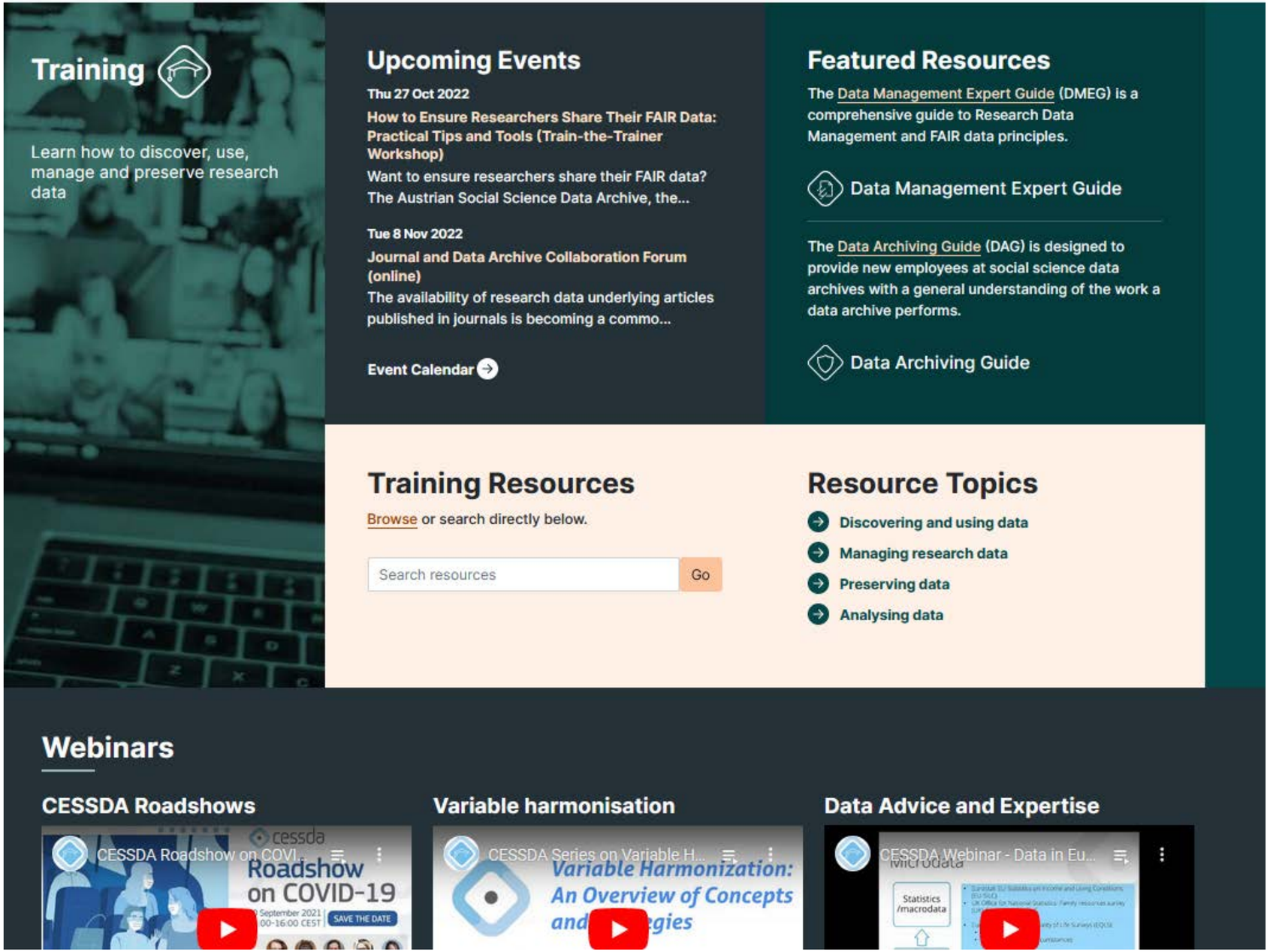
Metadata Validator

Validate metadata for compatibility with the CESSDA Data Catalogue and the European Question Bank.

https://datacatalogue.cessda.eu/

CESSDA Data Catalogue





https://www.cessda.eu/Training

CESSDA Data Management Expert Guide

- created by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (FAIR)
- 7 chapters following the data life-cycle

Data Management Expert Guide

https://zenodo.org/record/3820473#.YRZfs4hKhPY

dmeg.cessda.eu

Today’s Programme (CEST)

10.00 - 10.15 Welcome

10.15 - 10.50 Session 1: Open Science and data management planning basics: key concepts, tools and resources

10.50 - 11.00 Break

11.00 - 12.30 Session 2: Consent and ethical protocols in the Open Science landscape

12.30 - 13.30 Lunch break

13.30 - 13.45 Session 3: Data protection considerations and anonymisation in the GDPR context

13.45 - 14.30 Session 4: Licence frameworks @CESSDA archives

14.30 - 14.45 Break

14.45 - 15.15 Breakout rooms: Successes, challenges and potential in Open Science, Consent and Ethical considerations and Licences and Anonymisation training

15.15 - 15.30 Close

Main Objective of the workshop

1. Raise awareness of key tools and resources available for Open Science training

2. Enable a platform to exchange ideas regarding key training topics

3. Provide training materials and worksheets for future reuse

Next Session: Open Science and data management planning basics: key concepts, tools and resources

cessda.eu

@CESSDA_Data

Open Science and Data Management Planning: Key Concepts, Tools and Resources

Johana Chylíková, Ph.D.
Researcher, Data Manager, Czech Social Science Data Archive (CSDA)

27. October 2022

How to Ensure Researchers Share Their FAIR Data

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Overview

1. What is Open Science and why it is important?
2. Data Management Plan and why is it a must!

Perspective: How to explain the concepts of Open Science and Data Management to Researchers?

What is Open Science?

Open Science = Removing the barriers to share science (results, methods, theory…)

- Open Science According to OECD =
  "to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction" (OECD, 2015:7)

Open Science involves:

- **Open access to publications and research data**
- Open source software
- Open collaboration
- Open notebooks
- Open educational resources

Prerequisite of OS: digital technology that is broadly accessible

Closed science vs. Open science

- Current goal: Moving from the situation of the “closed science” towards the system where science is open to everyone who is interested
- Ultimate goal: Change of the system, science is open as possible

Why do we want Open Science?

Main reasons:

- Helps create high standards for scientific work (exchange of know-how and information)
- Driver for innovation (consequence of the fast exchange)
- Saves resources (time + money)

Why should we want it?

- Sociological perspective: Scientific knowledge is a product of social collaboration and its ownership belongs to the community
- Public perspective: Science must be made public so that citizens have access to it
- Democratic perspective: Access to knowledge must be equally distributed
- Economic perspective: Scientific outputs generated by public research are a public good that everyone should be able to use at no cost
- Infrastructural perspective: Efficient research needs available tools, data, applications etc.
- Pragmatic perspective: Production of knowledge will be more efficient if scientists work together

Institutional and individual actors of Open Science

Institutions:

- Universities, scholarly journals, project funders etc.

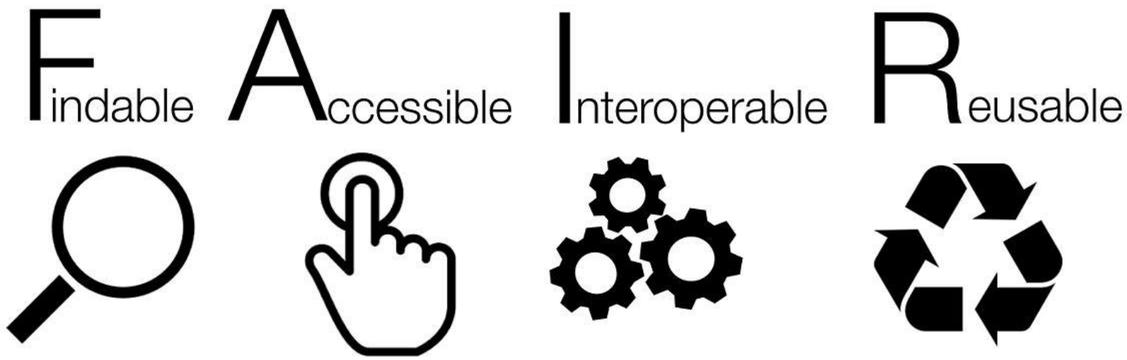
Individuals:

- Researchers, project managers, administrators etc.

How can an individual do open science in real life?

Be FAIR!

- What is FAIR? The well-know mantra =>



Being FAIR with Data and Metadata:

Findability

- Assigned PID
- Rich metadata
- Published metadata

Accessibility

- Retrievable by PID
- Metadata available forever (even after data are not available)

Interoperability

- Formal and broadly applicable language
- Use vocabularies
- References to other data

Reuse

- Rich metadata
- Licensed
- Meet domain standards

FAIR In the research cycle

Data collection:

- clarify usage rights
- cite the data and instruments involved in data collection

Processing:

- use open-source software

Storing the data:

- for a long-term preservation use a data repository
- attach persistent identifier to your data
- make sufficient metadata and publish them with open licence
- ensure links between publications, data and methods

Reuse:

- clear citation
- assign all necessary credits

Obstacles in Open Access to data

- Researchers feel the need to protect the data they produced with their own know-how
- Sometimes for understandable reasons:
  - Science as a competitive environment
  - Researchers building credit and prestige for themselves

How to deal with the clash of perspectives (Open Science vs. Being competitive in the field)?

=> Embargo on data sharing

Researcher (natural sciences) on a meeting for the project where it is mandatory to make data publicly accessible:

“These are data from experiments we will have worked for several years. We will not be able to have fully analyzed them in many years to follow. If we make them publicly available, someone else can take them and practically steal our work”.

The SSH Training Discovery Toolkit

- Directory of platforms for education and skills training

https://training-toolkit.sshopencloud.eu/

- Knowledge platforms, courses, videos, workshops, lessons
- For researchers, data archivists, students
- Most of 100 listed services/platforms in English
- Topics: Research data management, survey data, quanti/quali data analysis, software and programming, data visualisation
- Disciplines: social sciences, humanities, history, linguistics etc.

DMP Data Management Plan

Formal document that provides a framework for how to handle the data material during and after the research project.

But!

DMP is not only a FORMAL document required by the funder!

Researchers must understand that DMP is an extremely practical tool.

It is the job of archivists to convince researchers that:

- DMP is a must
- it is not just something that funders want
- DMP substantially improves the workflow in the project

Data Management Plan is:

- Practical
- Has considerable benefits
- Not difficult to make
- Allows for easy project management
- Makes data FAIRer

What researchers must know about DMP

DMP is a look into the future

When the project runs over several years,
the DMP is absolutely necessary.

With DMP researchers will know in advance:

- When the data will be collected and how
- How to handle the protection of respondents
- How to manage the data during and after the project
- Who will be responsible for what in the project
- The budget for everything related to data

A lot of diversity exists in DMPs because DMPs are built around the particular needs of the research project

= >

What researchers must know about DMP

- Researchers are sometimes uncertain about preparing the DMP
- Worried they would put in the DMP something wrong
- But they should not be afraid of being creative
- They should put in the DMP all the important information without worrying about doing something wrong
- They should bear in mind that the DMP is mainly for them to not get lost in their own project
- But should take into consideration requirements of the funder of the research project

Where is DMP strongly recommended or mandatory?

- Not all funders in European countries want the DMP
- Look into the DMEG, the country diversity in DMP
- see https://dmeg.cessda.eu/Data-Management-Expert-Guide/1.-Plan/European-diversity

DMP required:

- Horizon 2020
- DMP online via https://dmponline.dcc.ac.uk/

DMP in European countries

- In many European countries not mandatory yet
- It is recommended or mandatory by some funding institutions in the following countries:
  - Finland, Netherlands, Norway, Switzerland, UK

Recommended content of DMP for social sciences

- Following the DMP in DMEG
- dmeg.cessda.eu/
- DMP questions DMEG
- Editable version DM DMEG material – The handout for this presentation

1. Overview
2. Organising and documenting your data
3. Processing your data
4. Storing your data and metadata
5. Protecting your data
6. Archiving and publishing your data
7. Discovering data

DMP Content

1. Overview

- Title of the project, Description, Principal Researchers, Funder, Owner, Data Producer, Contacts, Roles in the project, Costs and resources

2. Organising and documenting data

- Data collection, Data organisation, Data type and size, File format, Folder structure and names, File structure and names, Documentation, Metadata, Metadata standard

3. Processing data

- Versioning, Interoperability, Data Quality

4. Storing data and metadata

- Storage, Backup, Security

5. Protecting data

- Ethical review, Informed consent, Personal data /confidential information, Intellectual property rights, Agreements with other stakeholders, Restrictions on data

6. Archiving and publishing your data

- Archiving, Data formats, Access

7. Discovering data

- Identification of needs, Search for data, Evaluation of data quality, Gaining access to data

References:

- Go Fair https://www.go-fair.org/fair-principles/
- CESSDA Data Management Expert Guide https://dmeg.cessda.eu/
- Foster Open Science https://www.fosteropenscience.eu/
- OECD (2015). Making Open Science a Reality. (OECD Science, Technology and Industry Policy Papers, 25). Paris: OECD Publishing.

THANK YOU!

johana.chylikova@soc.cas.cz

cessda.eu

@CESSDA_Data

Next on is a 10 minutes break, followed by “Consent and ethical protocols in an Open Science landscape”

cessda.eu

@CESSDA_Data

The Ethics of Consent Concerning Empirical Research and Data Archiving

Dimitri Prandner, MA PhD
Senior Scientist at Faculty of Social Sciences, Economics & Business of the Johannes Kepler University

Date: Oct. / 27 / 2022
Event: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Intro:
Why are consent and ethical questions central for empirical research and data archiving?

Dealing with the issue of consent is inherent to social scientific studies

- **When dealing with human beings, questions of ethics play a central role**
  - Very often this is shaped by normative assumptions and values

- **When it comes to data archiving further complications come into play**
  - F.A.I.R. data should be stored, made available and useable for scientific purposes, but there are considerations to be taken into account: unrestricted access to data could hurt individuals, could be misused or misinterpreted

- **Balancing act:**
  - As open as possible and as secure as necessary
  - Stakeholders and their roles in the process

Dealing with the issue of consent is inherent to social scientific studies

- Different laws and legal obligations of researchers, stakeholders and involved parties make not only a normative discussion
- Local legislation is often different and not uniform across Europe, some countries require ethics boards to approve each and every study, others have special rules for specific topics or disciplines, etc.
- Universities and research institutions handle it very differently

FAIR, but ethical? Two competing ideas when it comes to data archiving.

Preservation Policies

"Preservation Policies should be in line with the core mission of the organization as well as the legal framework in order to ensure permanent accessibility to archives' holdings or secure access to specific datasets (sensitive data, data under embargo period, etc.). Repositories need to determine the scope of preservation actions by identifying the collections to be preserved, their significance and the desired preservation period: how, for how long, and under which specific access conditions/restrictions."

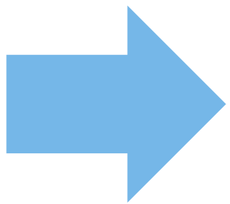
(Sensitive) Personal Data

"Data producers can deposit both, 'raw' (primary data that is not transformed in any way) and/or anonymized data. In practice, data archives often receive and accept 'raw' data. If it includes personal data, archivists usually work closely with data producers to make a decision regarding anonymization, since only anonymized data (i.e. data that do not contain any information that may lead to the identification of a particular individual) can be shared/distributed to secondary users. If the data contain too much personal information and the deletion of all this information would severely limit the scientific value of the data, the data are usually left as they are and the access to them is restricted.
This does not apply if data sharing is stated differently in the consent form, e.g. when respondents approved sharing their personal data."

Source: CESSDA Training Team (2022). *CESSDA Data Archiving Guide* version 1.0. Bergen, Norway: CESSDA ERIC.

What is expected from consent forms and how does this tie to ethics?

- An explanation of the purpose of research and the expected role of the research subject
- A description of the research topic and the type of data collected (especially sensible data)
- A description of any foreseeable risks a description of what will be done to minimize them (anonymization, pseudonymization)
- A description of how the data is processed, including who may have access to research data (relevant for archiving)
- An explanation of whom to contact for answers to pertinent questions about the research and the research subject's rights (including data protection offices)
- A statement that participation is voluntary



What is the role of an archive or the duty of data stewards?

What follows now:
An intro into creating consent forms and associated challenges

(1) Presentation by Walter Scholger:
The DARIAH ELDAH - Consent Form Wizard
a tool to create GDPR compliant consent forms
(2) Short Q & A of the opportunities such tools provide

Round Table on the Ethics of Data Archiving

cessda.eu

@CESSDA_Data

THANK YOU!

Dimitri Prandner@jku.at

cessda.eu

@CESSDA_Data

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Data protection considerations and anonymisation in the GDPR context

Cristina Magder / UK Data Service

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Training on open science: anonymisation

During your sessions consider including the following:

- Objectives of the session
- Context
- Key concepts
- Tools
- Exercises

Setting Objectives

Anonymisation in the context of data protection legislation can be a difficult and wide topic

When setting objectives always consider:

- Participants' level of understanding
- Focus of the session (quantitative/qualitative or both)
- Length of the session

Examples of Objectives

- Gain awareness of issues and considerations of data protection legislation and statistical disclosure for quantitative/qualitative data
- Gain basic knowledge of key concepts, techniques and tools in statistical disclosure control for quantitative/qualitative data
- (if tool introduced) Be able to implement semi-automated tools in your anonymisation process

Setting the Context

More and more research is now crossing country boundaries

=> leading to implications on data protection legislation

When setting the context of your training session try to provide practical examples on how different legislation applies (worksheet available on the shared drive)

Example of Data Protection Considerations

When personal information about people is collected or used in research data protection regulations apply such as:

EU GDPR (2018) & country specific laws

- a researcher based in a country from the European Union (country that adheres to EU GDPR) collects personal data about people living in a country from the European Union

e.g. DPA (2018) & the UK GDPR applies when

- a researcher based in the UK collects personal data about people anywhere in the world
- a researcher outside the UK collects personal data on UK citizens

e.g. DPA (2018), EU GDPR (2018) & the UK GDPR applies when

- a researcher based in the UK collects personal data about people across Europe

Examples of Key Concepts

Types of identifiers: direct (names, address etc.) vs indirect (education level, occupation, age etc.)

Types of disclosure: identity vs attribute

De-identification/Pseudonymisation: Process of removing or masking direct identifiers in personal data

Anonymisation: Process of ensuring that the risk of somebody being identified in the data is negligible; more than simply de-identifying the data, and often requires that indirect identifiers be further altered or masked

Types of anonymisation: Formal Anonymisation, Guaranteed Anonymisation , Statistical Anonymisation, Functional Anonymisation

Anonymisation and the DMEG

The CESSDA Data Management Expert Guide covers key anonymization topics including:

- Anonymisation methods
- Expert tips
- Case studies
- Anonymisation in practice

Qualitative Anonymisation Exercise Example

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.

Source: DMEG Anonymisation

Quantitative Anonymisation Exercise Example

Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier	Anonymisation method
Personal identification number	x			Remove
Full name	x			Remove/Change
Email address	x	x		Remove
Phone number		x		Remove
Postal code			x	Remove/Categorise
District/part of town			x	Categorise
Municipality of residence			x	Categorise

Create a dummy dataset and ask participants to identify direct and indirect identifiers, as well as methods to reduce the risk of disclosure

Source: DMEG Anonymisation

Further open access exercises on assessing disclosure control are available in Magder, Cristina. (2021, August 24). Facilitating Onwards Sharing of Safe and Clean Microdata [Train the Trainer Workshop]. Zenodo. https://doi.org/10.5281/zenodo.5566858

Introducing Tools in Your Training

A variety of useful open source tools are available for providing comparison between statistical disclosure control methods for quantitative data.

Tools should always come with the disclaimer that while quick and easy to explore what changes have biggest effect these can be more problematic when trying to define absolute risk i.e. the numbers might have no real meaning, and they can be time consuming and skill-intensive.

Example of Anonymisation Freeware

sdcMicro – R scripting + GUI, reference manual; new Shiny GUI – in detailed vignette, SDC methods and sdcMicro; in depth guide for SDC and sdcMicro

QAMyData – provides functionality for identifying direct identifiers with the use of regular expressions (RegEx)

Amnesia anonymisation tool – standalone tool, locally run with comprehensive documentation

µ-Argus – standalone tool recommended by Eurostat for government statisticians; software and manual

ARX – comprehensive open source software for anonymizing sensitive personal data; software and documentation

Main Takeaways

- Set clear target group and expected knowledge level from audience
- Set clear objectives and describe them when promoting your session
- Explain key concepts giving easily understandable examples and keep the audience engaged using mini exercises throughout the session (using Zoom polls, Menti exercises etc.)
- Allow time for questions and clarifications (ensure these can be submitted anonymously if the training is online)
- If introducing a tool chose the tool you are using and are most familiar with
- Always consider a short introduction to licences and how access levels is dependent on data contents

Next Session: Licence frameworks @CESSDA archives

cessda.eu

@CESSDA_Data

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Data protection considerations and anonymisation in the GDPR context

Cristina Magder / UK Data Service

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Training on open science: anonymisation

During your sessions consider including the following:

- Objectives of the session
- Context
- Key concepts
- Tools
- Exercises

Setting Objectives

Anonymisation in the context of data protection legislation can be a difficult and wide topic

When setting objectives always consider:

- Participants' level of understanding
- Focus of the session (quantitative/qualitative or both)
- Length of the session

Examples of Objectives

- Gain awareness of issues and considerations of data protection legislation and statistical disclosure for quantitative/qualitative data
- Gain basic knowledge of key concepts, techniques and tools in statistical disclosure control for quantitative/qualitative data
- (if tool introduced) Be able to implement semi-automated tools in your anonymisation process

Setting the Context

More and more research is now crossing country boundaries

=> leading to implications on data protection legislation

When setting the context of your training session try to provide practical examples on how different legislation applies (worksheet available on the shared drive)

Example of Data Protection Considerations

When personal information about people is collected or used in research data protection regulations apply such as:

EU GDPR (2018) & country specific laws

- a researcher based in a country from the European Union (country that adheres to EU GDPR) collects personal data about people living in a country from the European Union

e.g. DPA (2018) & the UK GDPR applies when

- a researcher based in the UK collects personal data about people anywhere in the world
- a researcher outside the UK collects personal data on UK citizens

e.g. DPA (2018), EU GDPR (2018) & the UK GDPR applies when

- a researcher based in the UK collects personal data about people across Europe

Examples of Key Concepts

Types of identifiers: direct (names, address etc.) vs indirect (education level, occupation, age etc.)

Types of disclosure: identity vs attribute

De-identification/Pseudonymisation: Process of removing or masking direct identifiers in personal data

Anonymisation: Process of ensuring that the risk of somebody being identified in the data is negligible; more than simply de-identifying the data, and often requires that indirect identifiers be further altered or masked

Types of anonymisation: Formal Anonymisation, Guaranteed Anonymisation , Statistical Anonymisation, Functional Anonymisation

Anonymisation and the DMEG

The CESSDA Data Management Expert Guide covers key anonymization topics including:

- Anonymisation methods
- Expert tips
- Case studies
- Anonymisation in practice

Qualitative Anonymisation Exercise Example

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.

Source: DMEG Anonymisation

Quantitative Anonymisation Exercise Example

Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier	Anonymisation method
Personal identification number	x			Remove
Full name	x			Remove/Change
Email address	x	x		Remove
Phone number		x		Remove
Postal code			x	Remove/Categorise
District/part of town			x	Categorise
Municipality of residence			x	Categorise

Create a dummy dataset and ask participants to identify direct and indirect identifiers, as well as methods to reduce the risk of disclosure

Source: DMEG Anonymisation

Further open access exercises on assessing disclosure control are available in Magder, Cristina. (2021, August 24). Facilitating Onwards Sharing of Safe and Clean Microdata [Train the Trainer Workshop]. Zenodo. https://doi.org/10.5281/zenodo.5566858

Introducing Tools in Your Training

A variety of useful open source tools are available for providing comparison between statistical disclosure control methods for quantitative data.

Tools should always come with the disclaimer that while quick and easy to explore what changes have biggest effect these can be more problematic when trying to define absolute risk i.e. the numbers might have no real meaning, and they can be time consuming and skill-intensive.

Example of Anonymisation Freeware

sdcMicro – R scripting + GUI, reference manual; new Shiny GUI – in detailed vignette, SDC methods and sdcMicro; in depth guide for SDC and sdcMicro

QAMyData – provides functionality for identifying direct identifiers with the use of regular expressions (RegEx)

Amnesia anonymisation tool – standalone tool, locally run with comprehensive documentation

µ-Argus – standalone tool recommended by Eurostat for government statisticians; software and manual

ARX – comprehensive open source software for anonymizing sensitive personal data; software and documentation

Main Takeaways

- Set clear target group and expected knowledge level from audience
- Set clear objectives and describe them when promoting your session
- Explain key concepts giving easily understandable examples and keep the audience engaged using mini exercises throughout the session (using Zoom polls, Menti exercises etc.)
- Allow time for questions and clarifications (ensure these can be submitted anonymously if the training is online)
- If introducing a tool chose the tool you are using and are most familiar with
- Always consider a short introduction to licences and how access levels is dependent on data contents

Next Session: Licence frameworks @CESSDA archives

cessda.eu

@CESSDA_Data

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Licence frameworks @CESSDA archives

Cristina Magder / UK Data Service

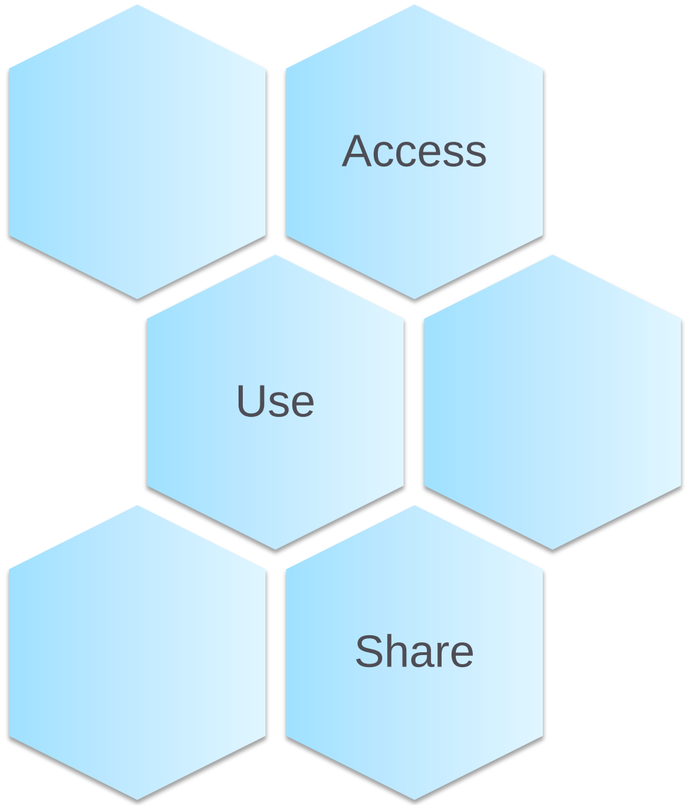
27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Encouraging Researchers to Licence Their Data

Always emphasise the role and importance of licences during training:



Open Licences Example: Creative Commons

Licence	Use and Share	Attribute the work (copyright)	Use commercially	Adapt/create derivatives	Change licence
CC0	✓	x	✓	✓	✓
CC BY	✓	✓	✓	✓	✓
CC BY-SA	✓	✓	✓	✓	x
CC BY-NC	✓	✓	x	✓	✓
CC BY-NC-SA	✓	✓	x	✓	x
CC BY-ND	✓	✓	✓	x	✓
CC BY-NC-ND	✓	✓	x	x	✓

Source - CESSDA DMEG: Licensing your data, inspired by Foter, 2015

Risk Evaluation and Data Licence

- Social sciences microdata risk assessment is a key factor in choosing the right licence.
- Introduce researchers to the concept of bespoke licences and access frameworks (used by data archives, national and institutional repositories to support further data access).
- Whenever possible provide concrete examples such as the ones covered shortly.

Licence Frameworks @CESSDA Archives

Presentation will follow shortly from:

- AUSSDA
- CROSSDA
- GESSIS
- UKDS

Licensing Models for Research Data @ AUSSDA

Open Access and Scientific Use

Dimitri Prandner, MA PhD
Senior Scientist at Faculty of Social Sciences, Economics & Business of the Johannes Kepler University

Date: Oct. / 27 / 2022
Event: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Guideline within AUSSDA:

As open as possible,
as secure as necessary!

- ☐ Societal responsibility

Open Access | OA

Scientific Use | SUF

Metadata is licensed via CC 0

Research data is licensed via CC BY.

Documentation is (typically) licensed via CC BY.

Datasets are meant to be used by students and educators (outside of tertiary education), policymakers, the interested public outside academia.

Datasets typically do not include sensitive data. Datasets are reduced, to meet GDPR demands and fulfill ethical consideration.

Research data is restricted, and academic credentials need to be provided for access (Account based).

Restrictions for manually controlled access can be put in place in special cases, when sensitive data is included.

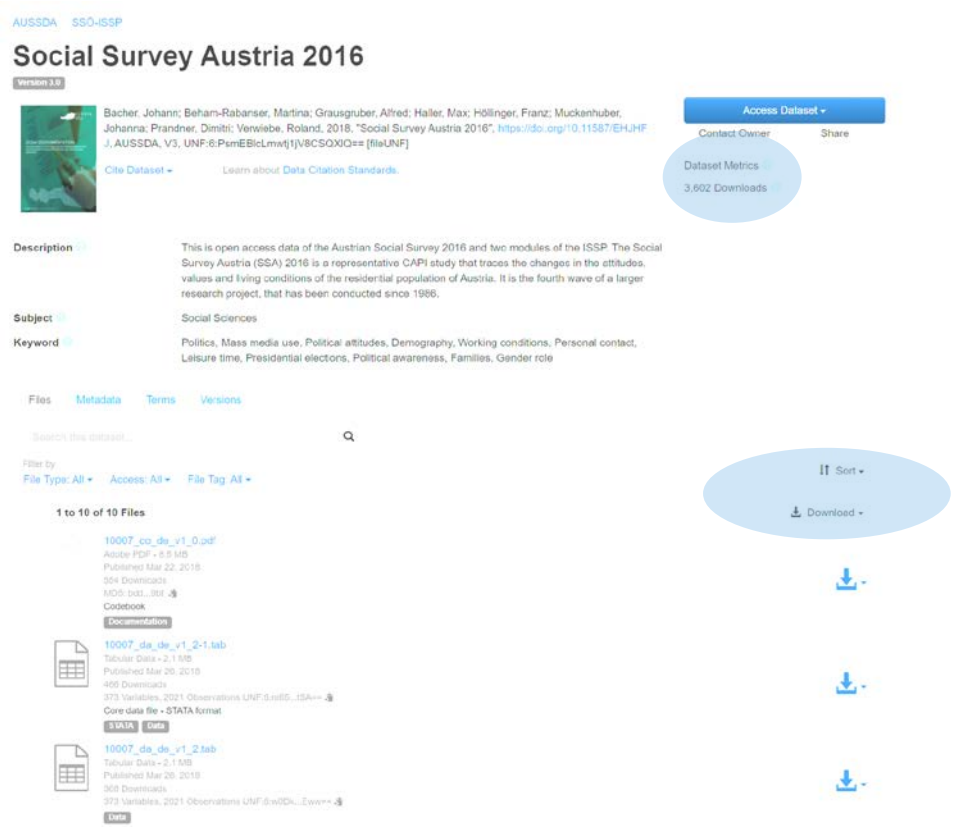
Documentation is typically licensed via CC BY.

Datasets are meant to be used for academic research (scientific purpose) as well as higher education and learning.

Datasets are GDPR compliant but take special regulations for scientific work into account.

Embargo

Open Access | OA



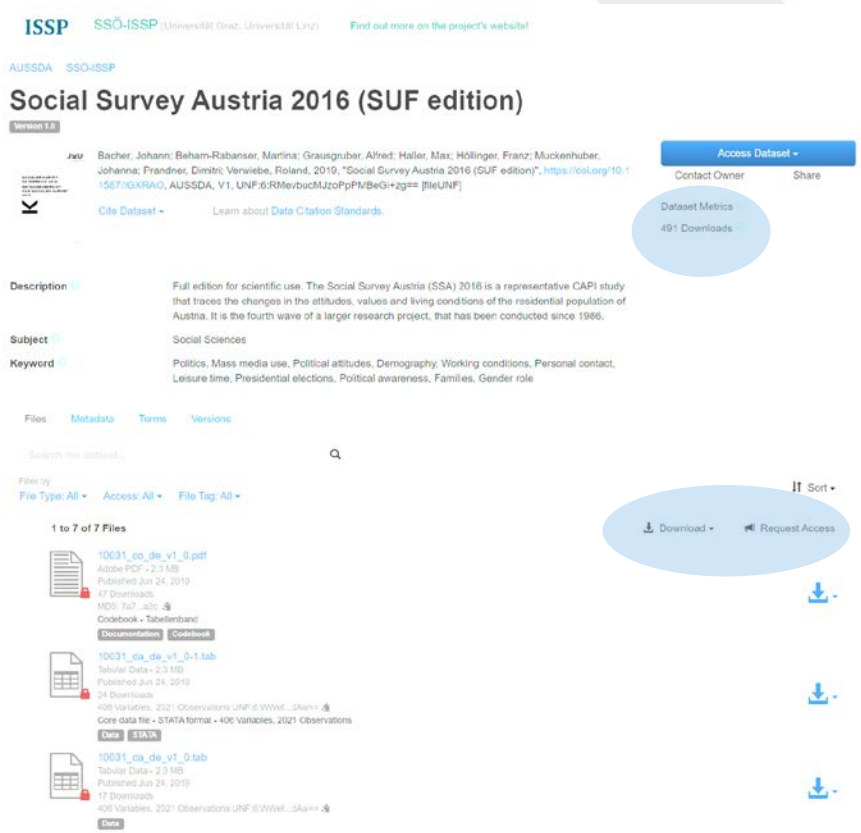
Reduced dataset,
some variables are not included,
others are recoded into broad categories

“Click to download”

Scientific Use | SUF

Nearly full dataset, all variables are included.

However, some information is slightly recoded to ensure protection of participants (e.g.: specific combination of job & sociodemographics)



Open Access | OA

Scientific Use | SUF

Metadata is licensed via CC 0

Research data is licensed via CC BY.

Documentation is (typically) licensed via CC BY.

Datasets are meant to be used by students and educators (outside of tertiary education), policymakers, the interested public outside academia.

Datasets typically do not include sensitive data. Datasets are reduced, to meet GDPR demands and fulfill ethical consideration.

Research data is restricted, and academic credentials need to be provided for access (Account based).

Restrictions for manually controlled access can be put in place in special cases, when sensitive data is included.

Documentation is typically licensed via CC BY.

Datasets are meant to be used for academic research (scientific purpose) as well as higher education and learning.

Datasets are GDPR compliant but take special regulations for scientific work into account.

Embargo

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Licence frameworks @CESSDA archives

Marijana Glavica / CROSSDA

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Croatian Social Science Data Archive (CROSSDA)

- Croatia become a member of CESSDA ERIC in 2019
  - Faculty of Humanities and Social Sciences, University of Zagreb (FFZG) was appointed by the Ministry of Science and Education to serve as a national coordinating institution for CESSDA ERIC
- CROSSDA was established in 2020 as a separate organisational unit at FFZG to serve as a Service Provider for CESSDA
- Dataverse is used as for data publishing

CROSSDA
THE CROATIAN SOCIAL SCIENCE DATA ARCHIVE

Licence Framework

A. Freely available data (Open Access)

- data is available under the terms of the Creative Commons Attribution 4.0 International license (CC BY 4.0)

no registration is required

B. Data available only for specific purposes

- for scientific use only
- for scientific use and teaching purposes

registration and accepting General Terms and Conditions for Data Use is required

Scientific purpose

- preparation and implementation of scientific research (projects, scientific papers, conference presentations, doctoral dissertations)
- postgraduate specialist theses
- theses at the level of graduate studies
- checking the results presented in the scientific work (reviewers or other researchers)

Why scientific use and teaching are separated?

- Croatia is a small country; representative samples can have about 1000 participants
- We can not be sure that undergraduate and graduate students fully acquired ethical norms
- If research is conducted on students - they could know each other - re-identification is possible not only by combining demographic variables, but also by knowing some answers

Lessons Learnt/Next Steps

- monitoring usage statistics is hard for Open Access data
  - no information about the purpose of data use
  - no information about who is using data (researchers, students, journalist, …)
- implementation of a "guestbook" is possible in Dataverse
  - it can be optional for Open Access data

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Licence frameworks @CESSDA archives

Oliver Watteler / GESIS – Leibniz Institute for the Scoial Sciences

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Introduction to GESIS

- Social science research data infrastructure
- Covers the entire life-cycle of empirical research (similar to ICPSR (USA) or FORS (CH)
- Is data collecting and data providing institution
- Offers a broad spectrum of research-based services for empirical social research
- “Archive”: Dept. Data Services for the Social Sciences (first founded in 1960)

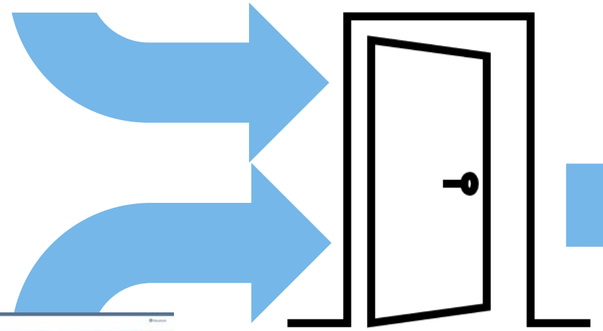
- Offers access to ~6,500 ‘studies’; mostly survey data
- Access via download off-site
- Secure data access on-site (Secure Data Center)
- Data Services:
  - Archiving BASIC / PLUS / PREMIUM
  - Research Data Management Training

gesis Leibniz Institute for the Social Sciences

Licence Framework

Different access classes

Archiving PLUS / PREMIUM (for cost)



Archiving BASIC (free of charge)

gesis

gesis Leibniz Institute for the Social Sciences

Usage regulations

Please note: Neither the depositor (individual(s), institute(s) etc.) nor GESIS bear any responsibility for the analysis, the methods used for the analysis, the interpretation with regard to contents of the data which is provided by GESIS.

For further information (e.g. on access categories, exact charges for special studies) please contact:

• Oliver Watteler, M.A., Tel. +49-221- 47694-412, mailto:oliver.watteler@gesis.org

You can find general advice concerning privacy protection on the GESIS website:

Usage regulations

research-based services for the social sciences. One of the tasks for the promotion of social science research is the archiving, documentation and long-term preservation of data, as well as the creation of user-friendly, high quality access to all information and data relevant to the social sciences (all these tasks together are called data services). The data services are provided by the GESIS Data Archive.

The GESIS Archive's holdings include the data depositor's original data and documents as well as if applicable additional material, which is the result of standardized documentation and processing at GESIS (e.g. adjusted data sets, variable reports etc.).

Data and documents are made available exclusively on the basis of these usage regulations.

1. General access conditions

Unless explicitly indicated differently, GESIS makes data and documents available only for scientific analysis carried out in a project limited to a specific time period in academic research and teaching.

Institutes and individuals outside academic research and teaching can apply for access in writing.

Data and documents are provided within the limits of access categories (see 2.).

cessda

Lessons Learnt and Next Steps

- Current setup of data ingest at GESIS too complex

- Setting up a “one door” (= single ingest point) for studies
  - Harmonize licenses for data deposition
- Planning an online portal for demanding access restricted studies
  - Harmonize licenses for data access
  - Remodel access classes

- Opening Secure Data Center for remote access

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Licence Framework @ CESSDA Archives

Cristina Magder / UK Data Service

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

UK Data Service

UK's largest collection of UK and international social sciences and population data

funded by the Economic and Social Research Council (ESRC)

provide users with access, support, guidance and training to facilitate high quality social and economic research and education

support the development of best practices for data preservation and sharing standards participants

UK Data Archive, the lead partner of the UK Data Service, a centre of excellence in acquiring, curating and providing access to the largest collection of social science and population data for over 50 years

Accessing Data

Open when possible, closed when necessary

We provide access to data via a three-tier licence/access framework:

- Open
- Safeguarded
- Controlled

UKDS Deposit Licence Agreement

At the UK Data Service, data are classified according to their level of detail, sensitivity and confidentiality.

All data depositors must sign a non-exclusive deposit licence agreement to have their data available via the UKDS.

The Agreement sets the responsibilities for both the data depositor and the UKDS.

The Collections Development and Data Curation teams work together to negotiate appropriate access levels for data.

UKDS End User Licence Agreement

All UKDS registered users must agree to the End User Licence Agreement.

The Agreement sets the responsibilities for both the registered user and the UKDS.

The conditions of use require researchers to act responsibly and ethically with the data.

In case of a breach the UKDS Licence Compliance Policy is followed.

UKDS Three Tier Licence Framework

Open

- No real disclosure risk – truly anonymous data or consent to share data as collected is in place. Under open licences; almost no restrictions on reuse

Safeguarded

- Zero to low real disclosure risk – effectively anonymised data. Requires authentication and authorisation e.g. registered user and End User Licence Agreement, special conditions might apply

Controlled

- Real disclosure risk – de-identified data (personal data) and/or highly sensitive. Requires project approval, user vetting and training; access via a safe setting; output checking

UKDS Three Tier Access Framework

Open Access

- Registration not required

Safeguarded Access

- Registration required
- Special conditions might apply

Controlled Access

- Registration required
- Access available via the Five Safes Framework

Five Safes Framework

Safe data: treated and de-identified
Safe projects: approved by data owners/Research Accreditation Panel and must demonstrate public good
Safe people: complete the Safe Researcher Training and gain Accredited Researcher status
Safe settings: environment prevents unauthorised access i.e. UKDS SecureLab
Safe outputs: screened and approved to ensure no disclosive information leaves the Safe Settings

Thank You!

cessda.eu

@CESSDA_Data

Train the Trainer Workshop: How to Ensure Researchers Share Their FAIR Data: Practical Tips and Tools

Close

Cristina Magder / UK Data Service

27 October 2022

cessda.eu @CESSDA_Data

DOI: 10.5281/zenodo.7252952

Recap and Further Steps

- Overview of today
- All materials will be available via Zenodo (with a DOI)
- Recordings will be available on CESSDA Training YouTube Channel
- Participants will receive a follow-up email containing all relevant links

Thank you all!

We'll be in touch soon.

cessda.eu

@CESSDA_Data