1. Data discovery and the EUI Library Data Portal
2. Data protection, database copyright and ethical use
3. Data management plans
4. Managing data during the research project cycle
5. Repositing and preserving research data
6. Research data management in EU Horizon 2020 projects
7. Open Data
8. Qualitative data in the humanities and social sciences
9. Infrastructure, software and support
10. International research data guidelines

This EUI Library Guide is issued in support of the research lifecycle, encompassing data discovery; data management planning; data use; data protection; data preservation and open data.

1. DATA DISCOVERY AND THE EUI LIBRARY DATA PORTAL

This section describes access to EUI licensed data resources; access to data hosted by other institutions; and access to open data resources.

1(a) The EUI Library Data Portal

The Library maintains an indexed Data Portal\(^1\) which (i) provides access to licensed resources for EUI members and (ii) introduces open data resources of relevance to EUI research themes. All resources in the Data Portal have an online guide with:

- Data description and scope
- Time-period coverage and release/wave information
- Support links (online manuals, software transfer routines, user networks)
- Terms and conditions of access and use.

Data provided to EUI members under Library license are accessible by internet protocol (IP) proxy, or via the Library’s restricted data server for micro-socioeconomic data. Modes of access for every resource are described in the Data Portal. There are four sub-directories:

- Macro data directory providing access to national, regional and global economic, political and historical resources\(^2\)
- Micro data directory providing access to household, family, individual and company-level resources\(^3\)
- European, EU and Euro Area directory providing data for research on pan-European topics, EU states, European sub-state regions\(^4\)
- Data Topics directory helping match research topics with data resources.\(^5\)

\(^2\) https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/MacroDataSet
\(^3\) https://www.eui.eu/Research/Library/RequestForms/Register-micro-data
\(^4\) https://www.eui.eu/Research/Library/ResearchGuides/Economics/EuropeanEUandEurozoneStatistics
1(b) Access to data at other institutions

EUI members who require access to restricted data at another facility should apply for access early in their research projects because application periods can be lengthy. Some sociological, qualitative and micro geo-coded data can only be accessed at issuers’ secure, on-site facilities. Contact the Library for assistance with access applications.6

EUI members who require access to unpublished datasets (eg. underlying data associated with a publication) should contact the Library before writing to data creators/owners. In some instances, it may be possible for EUI members to obtain access via library consortia.

1(c) Access to open data resources

Open data refers to the trend among scholars, government agencies and international organisations to share datasets, codebooks and software via the internet.7 Open research datasets can be located via the re3data8 registry of research data repositories, which is indexed by discipline, sub-discipline, data type and host location; and via Google Dataset Search.9

Open data is further discussed in Section 7.

2. DATA PROTECTION, DATABASE COPYRIGHT AND ETHICAL USE

This section provides an overview of data protection and ethical data use; database copyright; and individual data user undertakings. Further information is provided in the EUI Guide to Good Data Protection Practice in Research.10

2(a) Data protection

Special terms and conditions apply to access and use of qualitative and micro-socioeconomic data. This reflects the sensitive nature of observations about human subjects, families and households. Such terms and conditions apply to (i) data gathered and used during research projects, (ii) data hosted by the Library for EUI members and (iii) data provided directly to EUI members under third-party license. Terms and conditions of access and use for every dataset hosted by the Library, are provided in the ‘full details’ section of the individual resource guides in the Library’s Data Portal.11

Persons, families and households cannot be identifiable in any dataset. When using micro datasets, users must not attempt to identify any individual, family or household.

The storage, transfer and use of personal data is subject to data protection rules.12 The processing of particularly sensitive data (eg. pertaining to ethnicity, health, religion,
political or sexual orientation, and biometrics) requires even stricter safeguards. Those rules are currently under revision in light of the new European Union General Data Protection Regulation (GDPR) which came into effect on 25 May 2018 and the new rules on data protection concerning EU institutions and agencies. As stated in the EUI Guide on Good Data Protection Practice: “As a general rule, the processing of such personal data is prohibited. However, Article 7 of the EUI’s Data Protection Policy does allow it to be lawfully processed in specific circumstances. The most common in research is upon the data subject’s explicit consent.”

When collecting (eg. via interviews or forms), storing, using or transferring personal data; scholars are responsible for obtaining the specific informed consent of subjects (Section 4(e) below). Scholars are also responsible for preserving the confidentiality of data observations pertaining to human subjects, families and households. Consent must be intelligible: it should refer clearly and precisely to the scope and the consequences of the data processing, and the retention period. Blanket consent without specifying the exact purpose of the processing is not acceptable; it must also be documented by the researcher.

EUI members must use the online micro data access form when applying for access to micro-socioeconomic data hosted on the Library’s restricted server. A separate form is required for every dataset required. Applicants must also sign (i) the EUI ‘Terms and Conditions of Use of Micro Data’ paper form, and (ii) the data provider’s terms and conditions’ paper form. Both forms can be signed at the Badia Library (085) or the Economics Departmental Library (Villa La Fonte, 035). Some providers (eg. Eurostat) require the submission of a detailed project proposal, and parallel registration, before access is granted.

There is no off-campus access, VPN access, or laptop access to micro-socioeconomic data hosted by the EUI Library. Non-EUI members and short-term visitors do not have access to EUI-hosted micro data. Further information is provided in the EUI Guide to Good Data Protection Practice in Research and the OECD’s guidelines on Research Ethics and New Forms of Data for Social and Economic Research.

2(b) License agreements and database copyright

Access to, and use of, databases provided by the EUI Library are subject to contractual license agreements and copyright terms. Full details are on the Library’s Terms and Conditions’ web page. Data users are individually responsible for compliance with terms and conditions of access and use. Violation of license terms puts at risk other EUI members’ future access to data resources. All EUI users must scrupulously abide by the terms and conditions of access to, and use of, data provided under license to the EUI community.

---

15 [https://www.eui.eu/Research/Library/RequestForms/Register-micro-data](https://www.eui.eu/Research/Library/RequestForms/Register-micro-data)
18 [http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx](http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx)
Users may not distribute or allow any other party to have access to data which is provided under license; users may not modify or create a derivative work of the licensed materials without the permission of the licensor; users may not remove, obscure or modify any copyright or other proprietary notices included in licensed materials; users may not use licensed materials for commercial purposes; users may not retain or distribute substantial portions of a database, and must comply with any post-project data destruction undertakings in the license.19

2(c) Agreements between individual users and data providers

Some data issuers require that access contracts be established directly with end-users. If a data issuer also requests the counter-signature of an EUI administrator (‘guarantor’), contact the Library. Upon completion of the access application, the Library will request that the EUI data user sign an internal undertaking to abide by the terms and conditions of access and use.20

3. DATA MANAGEMENT PLANS (DMPs)

Data management plans are short documents, normally required by science funding agencies. Research data management is carried out by scholars throughout the duration of a research project. Data management plans should address:

- How data is generated and/or sourced
- How data is used, elaborated and organised
- How data, and data subjects, are protected
- How data, code and ancillary elements are described and documented
- How data is stored and secured, and how long it will be retained
- How data authorship and credit are assigned
- How data is preserved
- How, whether, and under what terms, research data outputs can be shared.

During research projects it is important to keep an accurate record of changes to dataset design, folders, files, variables and versioning. Scholars undertaking experiments should ensure that pre-agreed terms of disclosure are documented and preserved. The data and methodology should be clearly documented. Where applicable, code should be preserved for repositor and sharing. Metadata – ‘data about data’ – should be carefully assigned (Section 5(c) below).

Data management plans (DMPs) can be used as the basis for determining whether, when, how, where and under what terms, research data outputs can be openly shared – or shared under more restrictive terms and conditions. Open data refers to the trend among scholars, government agencies and international organisations to share data outputs and ancillary materials via the internet. The FAIR data principles (findable,
accessible, interoperable and re-usable) must be considered during the preparation, and revision, of data management plans.21

EUI members who are required to submit a data management plan – either as part of a funding proposal or during a research project – should contact the EUI Library for assistance.22 Tools such as DMPonline23 – the Digital Curation Centre’s data management planning tool – can be used to prepare a structured data management plan, complying with (eg.) EU Horizon 2020 requirements. To use DMPonline, enter an email address, name of organisation and create a password. EUI users should select ‘other organisation’ from the drop-down menu. First-time users are taken to the ‘edit profile’ section of the DMPonline platform. The Principal Investigator (P.I.) should be identified in the data management plan. For research teams, the P.I. can assign co-author rights to the data management plan, by entering email addresses and assigning status as ‘co-owner’, ‘editor’ or ‘read only.’

The Principal Investigator should be the contact person for decisions regarding whether, when, how, where and under what terms, research data outputs might be openly shared. If the Principal Investigator is not the same person as the project Data Manager, this should be stated. In international collaborative projects, the name of the person who has final authority regarding decisions on the sharing of data outputs should be indicated. Unless otherwise stated, the Principal Investigator undertakes this role.

Science funders normally require a revised DMP at the mid-point of the research project and a definitive DMP as a project deliverable. It is important to update plans throughout the research project by incorporating changes to the scope of the project; new data generation and use; and changes in the composition of the research team or consortium.

4. MANAGING DATA DURING THE RESEARCH PROJECT CYCLE

Data should be carefully managed throughout the duration of the research project. Particular attention should be given to data input; quality control; dataset design; folder structure; file structure; file naming; variable naming; software format; documentation and codebooks; security and backup during the project; data protection; informed consent of data subjects;24 anonymisation;25 copyright; preservation; archiving; data citation and – where possible – sharing data outputs as open data. The following guides are particularly useful for scholars working in the social sciences and humanities: Research Data Management Toolkit by JISC (2018)26 Expert Tour Guide on Data Management by the Consortium of European Social Science Data Archives (2017)27 Research Data Lifecycle by the UK Data Service (2017)28 and Guide to Social Science Data Preparation and

---

22 econlibrary@eui.eu
23 http://www.dcc.ac.uk/dmponline
26 https://rdmtoolkit.jisc.ac.uk/
27 www.cessda.eu/DMGuide
28 https://www.ukdataservice.ac.uk/manage-data/lifecycle
Archiving by the Inter-university Consortium for Political and Social Research (2012).  

4(a) Data input and quality control

The generation and/or sourcing of data should be informed by high-quality control standards. Datasets which are generated from empirical work during a research project should adhere to the ethical best-practices of the discipline or sub-discipline. In cases where research datasets are partially based on pre-existing resources, scholars should ensure that the underlying sources are of high quality, and that there are no infringements of copyright. Contact the Library for assistance. The ingestion of observations into a research dataset should be undertaken with close attention to accuracy, completeness and consistency. Data codes and variables should be assigned in a logical, accurate and sustainable manner.

4(b) Folders, files, variables, format and versioning

The design of research datasets should be carefully considered at the outset of the project. Dataset design varies by conventions of disciplines and sub-disciplines, medium of data, types of variables, units of analysis, methodology, relationship between data elements, and whether or not the dataset is part of a series. Clear and consistent metadata for folders, files, variables and versioning helps facilitate future data retrieval, reuse and replicability.

The folder structure of the research dataset should be considered at the beginning of the project (e.g. hierarchical / horizontal). Qualitative datasets containing text, interviews, images &c. may require individual files for every element. File names should be standardised and consistent, e.g. date, descriptor, version. Variables, such as age, country and sex should be clearly tagged, avoiding special characters and spaces. Temporary identifiers should be removed from the schema.  

Files should be systematically named, using a standardised date system (YYYY-MM-DD) or other consistent versioning.

The software format of the dataset should facilitate flexible use of the data. Scholars using one format during a research project, may consider a different format for preservation – taking into consideration accessibility via open source software (if the dataset can be shared as open data). Details of how to submit datasets to the EUI Research Data collection in the Cadmus repository are in Section 5.

4(c) Documentation and codebooks

Clear and accurate documentation should be provided about the purpose, context and methodology of the research project; and about the research data output. Good documentation makes datasets findable, accessible, interoperable and re-usable (FAIR data principles).  

Documentation should include a detailed description of the dataset, providing information about folders, files, variables, versioning, and – where applicable –

29 https://www.icpsr.umich.edu/files/deposit/dataprep.pdf
30 https://www.ukdataservice.ac.uk/manage-data/format/organising
information about problematic values, missing observations and weightings. Codebooks, questionnaires and data dictionaries should be included. A concise note on methodology/methodologies, should be provided, along with information on how and when the data was generated/collected; how the data was collated and elaborated during the research project; and how the dataset might be used for future research.

4(d) Security and backup during the research project

During research projects it is important to keep data secure at all times. Scholars should use a desktop computer for data elaboration, and make regular backups on the EUI network server or on a secured external memory device. In accordance with contractual agreements, micro-socioeconomic data hosted by the EUI Library can only be accessed and elaborated on desktop computers. Preliminary findings and associated documentation should be kept in locked storage when not in use. The EUI ICT Service maintains a strong password policy to prevent the loss, exposure, or corruption of sensitive information.

4(e) Data protection, informed consent of data subjects and anonymisation

The data protection requirements for data access, described in Section 2(a) above, also apply to data gathering and use throughout the research cycle. Further information is provided in the EUI Guide to Good Data Protection Practice in Research. The mode of consent obtained from subjects depends on the nature of the research project, the kind of data collected, and how and when the data will be used. Where possible, it is recommended to obtain written consent, using a template appropriate to the discipline or sub-discipline in which the research is being conducted.

Dataset creators are responsible for the anonymisation of sensitive data observations. Anonymisation techniques include: data masking (partial data removal and data quarantining); pseudonymisation; aggregation (cell suppression, inference control, rounding, sampling &c.); and banding. It is not possible to publish a dataset containing significant portions of data sourced from pre-existing copyrighted databases governed by contractual license. The Library provides advice on data protection and copyright.

5. REPOSITING AND PRESERVING RESEARCH DATA

EUI members can preserve and share their research data outputs by submitting them for reposi in the EUI Research Data collection of the Cadmus repository. This section of the Guide provides information on preparing data for long-term preservation and, where possible, sharing as open data. Section 5(c) below, explains how to create metadata – which are essential for reposi, retrieval and reuse.

32 https://www.icpsr.umich.edu/files/deposit/dataprep.pdf
33 https://www.eui.eu/ServicesAndAdmin/ComputingService/PolicyDocuments/StrongPasswordPolicy
35 Anonymisation techniques are described in in the Anonymisation Code of Practice, UK Information Commissioner’s Office (2012, Appendix 2): https://ico.org.uk/media/1061/anonymisation-code.pdf The UKDS provides a guide to anonymisation for both quantitative and qualitative data at: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation
36 econlibrary@eui.eu
37 https://cadmus.eui.eu/handle/1814/64544
5(a) Preparing data for reposít

EUI members who wish to submit a dataset for inclusion in the EUI Research Data collection of the Cadmus repository should first complete the online submission form. Library staff will use the information from the form to generate metadata in the Dublin Core schema. An appointment for data transfer will be made by Library staff.

When preparing submissions, it is advisable to create two dossiers: one for data and one for documentation. Large datasets should be submitted in archive format (eg. .zip). Data should be submitted in original file format version. Subsets must be accommodated within the folder structure – not as multiple repository entries. New entries can be created for subsequent iterations of the dataset. Documentation should include a concise overview of the research project and methodology. Codebooks and other relevant documentation should be provided in PDF/A format, for preservation in an adjacent ‘Docs’ folder.

Scholars presenting datasets for inclusion in the EUI Research Data collection of the Cadmus repository should pay particular attention to data quality control, dataset structure and data protection. It is not possible to reposít unstructured data elements and observations in the repository.

5(b) Completing the submission form

EUI members wishing to reposít data, should complete the online submission form. By submitting the form, EUI members acknowledge that the dataset for reposít is the result of original data generation; or is the output of significant, value-added, elaboration of pre-existing sources. Datasets presented for inclusion in the repository must be the output of research by a current EUI member – or at least one EUI member of a research team. The name of the Principal Investigator, researcher(s), and – where applicable – technical collaborator(s), must be provided. EUI email contacts must be given. If the project is undertaken in the context of a consortium, the name of the Data Manager should be provided (if different from the Principal Investigator).

By completing the submission form, EUI dataset creators certify that their work complies with the Code of Ethics in Academic Research of the European University Institute. The source(s) of the data must be indicated. If the dataset is the output of original data generation, details must be provided. If the dataset is derived from pre-existing sources, those sources must be clearly indicated (eg. data creator, institutional source and/or publisher).

EUI members submitting data to the repository should state whether or not the dataset can be shared as open data or be subject to embargo. The Library can offer guidance on data protection and database copyright. In some cases it may be possible to create a

---

38 https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
39 http://dublincore.org/metadata-basics/
40 https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
limited public version of a larger restricted dataset. Embargo status can change over time.

Persons, families and households cannot be identifiable in any dataset (Section 4(e) above). The collection and use of observations relating to ethnicity, health, religion, orientation, biometrics &c. are subject to data protection laws. Creators of research data outputs which have been elaborated from pre-existing copyrighted sources may need to obtain the permission of rights' owners before open data sharing. It is not possible to publish a dataset containing substantial portions of data sourced from pre-existing databases governed by contractual license. The EUI Library can provide advice, and can also assist scholars to reposit datasets in discipline-specific data repositories and the multi-disciplinary Zenodo repository.42 Major data repositories are indexed in the international re3data registry43 and Google Dataset Search.44

5(c) Metadata

Metadata are 'data about data' presented in a systematic schema. Accurate metadata are necessary for the organisation, use, repositing, sharing and machine-discovery of datasets. Throughout the research cycle, it is important to keep a detailed and updated record of data capture, input, use and elaboration. An introduction to metadata standards for social science and humanities’ data is available from the Digital Curation Centre.45

Metadata elements can be used as a ‘checklist’ to determine whether, when, how, where and under what terms, research data outputs can be shared as open data. Metadata should be consistent throughout the research project. Some research data outputs may require multi-lingual metadata.

EUI members submitting a dataset for inclusion in the EUI Research Data collection of the Cadmus repository should first complete the Library’s online submission form.46 These are the principal metadata fields:

NAME(S) OF DATASET CREATOR(S)
The name, or names, of the scholars and technical collaborators who created the dataset must be provided. The name of the Principal Investigator must be given if the dataset has been created by a research team. If the project is undertaken in the context of a consortium, the name of the Data Manager must be provided (if different from the P.I.). Where researcher IDs are available, eg. ORCID, these should be provided. The CRediT framework provides guidelines on authorship and credit for research outputs.47

EMAIL CONTACT(S)
The EUI email address of the dataset creator(s) must be provided.

42 http://zenodo.org/about
43 http://www.re3data.org/
44 https://toolbox.google.com/datasetsearch
45 http://www.dcc.ac.uk/resources/subject-areas/social-science-humanities
46 https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
47 http://casrai.org/credit
TITLE OF DATASET
The title should succinctly convey the nature and scope of the dataset.

DESCRIPTION OF DATA
A meaningful abstract, describing the data, the purpose of the research project and the methodology, must be provided. The Library can help edit abstracts.

SOURCE(S) OF DATA
The source(s) of the data must be clearly indicated. If the dataset has been *generated during a research project*, this should be indicated with details of data collection methods (eg. survey parameters). If the dataset is *derived from a pre-existing database*, all source(s) must be clearly cited; eg. publisher, issuer. If there are multiple sources, all must be cited.

TYPE OF DATA
The type of data must be indicated: eg: statistical; textual; observational; computational; experimental; simulational &c.

YEAR OF COMPLETION OF DATASET
The date of completion of the dataset must be provided. If part of a data series/wave, this should be indicated.

DATE-RANGE COVERAGE OF DATASET
The start- and end-dates of dataset coverage must be provided.

GEOGRAPHICAL COVERAGE OF DATASET
Where applicable, the geographical scope of the dataset (national, regional, global &c.) should be indicated.

FORMAT OF DATA
The software format and version must be given (eg. Stata 15, .csv, Excel, .txt, &c.).

CODEBOOK / SUPPORTING DOCUMENTATION
Codebooks and supporting documentation should be provided in PDF/A format.

ACCESS STATUS
The status of access to the data must be indicated. The status ‘open data’ should be assigned to datasets that are to be made publicly available via the internet. If the data is subject to embargo, the expiry date of the embargo should be indicated, where applicable.

LICENSE
Submitters should chose a license option from the drop-down menu: CC-BY (international)48 or CC0.49

DIGITAL IDENTIFIER
The Library will assign a unique object identifier to the dataset, for the purpose of locating, linking and citation.

48 [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)
49 [https://creativecommons.org/share-your-work/public-domain/cc0/](https://creativecommons.org/share-your-work/public-domain/cc0/)
LANGUAGE
Where applicable, multi-lingual documentation, tags, questionnaires and variable descriptions should be provided.

FUNDING STATEMENT
The name of funding bodies, and research grant numbers, should be provided where applicable.

RELATED PUBLICATIONS
Bibliographical details of publications based on the dataset, if any, should be listed with links to abstracts and, where possible, full-texts.

PROJECTED FUTURE WAVES OF DATASET
In cases where it is intended to generate future iterations of the dataset, details should be provided.

DATASET CITATION
A short 'ready-to-use' data citation should be provided, incorporating core descriptive elements: family name; first name; year; dataset title; ‘EUI Cadmus Research Data’. For example: ‘Steinmo, Sven; D’Attoma, John; Noyon, Sanne Maria; Zhang, Nan (2018): Willing to Pay? Dataset. EUI Cadmus Research Data.’

Data sources should be cited in EUI theses, working papers, books and other publications (eg. by using the citation format above). When submitting details of research publications to the EUI Cadmus repository, scholars should cite datasets which support the research findings. Further information about data citation is provided by DataCite and the Data Citation Synthesis Group.

6. RESEARCH DATA MANAGEMENT IN EU HORIZON 2020 PROJECTS

EUI project managers who are preparing EU Horizon 2020 applications are required to submit preliminary information about data management in the proposed project. This information is evaluated by the European Commission under the criterion 'impact.' When completing the general information section of the Horizon 2020 application, EUI project applicants should address the following questions: (i) What types of data will the project generate/collect? (ii) What standards will be used? (iii) How will the data be exploited and/or shared for verification and re-use? (iv) If data cannot be made available, explain why (v) How will the data be curated and preserved? The EUI Library assists project managers with the data management sections of funding applications and data management plans.

When a project proposal is approved, a data management plan (DMP) must be provided within six months of the start of contract. The European Commission mandates two

50 http://cadmus.eui.eu/
51 Datacite metadata schema: https://schema.datacite.org/
52 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego, CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
53 https://ec.europa.eu/programmes/horizon2020
55 econlibrary@eui.eu
further versions of the DMP; one at mid-point and one as a final project deliverable. The EC provides further details in the Guidelines on Data Management in Horizon 2020.\(^{56}\) The DMPonline tool\(^{57}\) (Section 3 above) can be used for Horizon 2020 data management plans, by selecting the Horizon 2020 template in the Funder section.

The European Commission’s policy is for research data to be “as open as possible, as closed as necessary”.\(^{58}\) A list of opt-outs is provided by the EC.\(^{59}\) The Horizon 2020 research data pilot requires grant beneficiaries to deposit research data outputs in a research data repository. The EC does not recommend specific repositories. The guidelines state that “as far as possible, projects must then take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data. One straightforward and effective way of doing this is to attach Creative Commons Licences (CC BY or CC0) to the data deposited.”\(^{60}\)

6(a) European Open Science Cloud (EOSC)

The European Open Science Cloud portal was launched in November 2018.\(^{61}\) The EOSC portal enables the discovery of data; provides tools for data analysis; and gives information about storage, computation, training and security. Content and services are available via the EOSC catalogue.\(^{62}\) To mark the launch of the EOSC Portal, the European Commission published two supporting documents: Prompting EOSC in Practice\(^{63}\) which explains the governance of the new service, and Turning FAIR into Reality\(^{64}\) which explains how to make research data ‘Findable, Accessible, Interoperable and Reusable.’ The European University Institute endorsed the European Open Science Cloud Declaration in 2017.

6(b) Social Sciences and Humanities Open Cloud (SSHOC)

The Social Sciences and Humanities Open Cloud project was launched in January 2019. SSHOC will constitute the social sciences and humanities cluster of the European Open Science Cloud. SSHOC has been established “as part of the European Union’s Horizon 2020 research and innovation programme… It aims to bring together existing and new infrastructures from social science and humanities ERICs (European Research

\(^{57}\) http://www.dcc.ac.uk/dmponline
\(^{61}\) https://eosc-portal.eu/
\(^{62}\) https://catalogue.eosc-portal.eu/home
Infrastructure Consortiums) and foster interdisciplinary research and collaboration.\textsuperscript{65} 47 partners are involved, under the coordination of the Consortium of European Social Science Data Archives (CESSDA).

7. OPEN DATA

There is a growing trend among scholars, government agencies and international organisations to share data and associated documentation, code, software and tools. Open data resources are available via the internet. Major data repositories are indexed in the international re3data registry\textsuperscript{66} and Google Dataset Search.\textsuperscript{67}

By carefully generating the metadata elements listed in Section 5(c) above, scholars will have a ready checklist for determining whether, when, how, where and under what terms, research data outputs can be shared as open data. Not all research data outputs can be openly shared. The two most significant considerations when determining whether a research dataset can be made available on an open data basis relate to data protection (Section 2(a)) and database copyright (Section 2(b)). The EUI Guide to Good Data Protection Practice in Research provides further information.\textsuperscript{68}

Data access status may change over time. Data can be made openly available for all users via the internet; data can be subject to pre-access registration terms; data can be subject to user contract (sometimes requiring a project proposal); data can be embargoed for a defined period (or indefinitely) and data can be restricted to on-site access and use. In some cases it may be possible to create a limited public version of a larger, restricted dataset. Data can also be repositoned solely for preservation purposes (dark archive).

8. QUALITATIVE DATA IN THE HUMANITIES AND SOCIAL SCIENCES

This section treats access to, and use of, qualitative data in the humanities and social sciences. Examples of qualitative data include; minable text, interview transcripts; images; audio and video recordings; survey diaries; archival material; field notes, and free-text surveys. The definition of ‘data’ varies across academic disciplines and sub-disciplines – especially where there is a mix of qualitative and quantitative methods. It is important that dataset design be located in the culture of the discipline or sub-discipline in which the research is undertaken.

8(a) Access and terms and conditions of use: qualitative data

The generation, handling, use and sharing of qualitative data in the social sciences and humanities is subject to strong ethical considerations and standards. Qualitative data can be generated (eg.) from surveys, free-text responses to interview questions, focus group recordings or experimental simulations.

\textsuperscript{65} https://www.cessda.eu/About/Projects/Current-projects/SSHOC
\textsuperscript{66} http://www.re3data.org/
\textsuperscript{67} https://toolbox.google.com/datasetsearch
In all cases, subjects should be informed of their rights as established by jurisdictional data protection legislation, and best-practice guidelines from scholarly societies in the relevant discipline or sub-discipline. When collecting (eg. via interviews or forms), storing, using or transferring personal data, scholars are responsible for obtaining the specific informed consent of subjects. Scholars are also responsible for preserving the confidentiality of data observations pertaining to human subjects, families and households. Consent must be intelligible: it should refer clearly and precisely to the scope and the consequences of the data processing, and the retention period. Blanket consent without specifying the exact purpose of the processing is not acceptable; it must also be documented by the researcher. Due to the sensitive personal nature of many qualitative data observations, scholars should pay particular attention to ethical standards when handling data relating to ethnicity, health, religion, orientation, biometrics &c. Persons, families and households cannot be identifiable in any dataset. The linking of variables on ethnicity &c. to individuals, families or households, is governed by data protection legislation and academic best-practice.

When accessing pre-existing qualitative data, it is important for scholars to familiarise themselves with the terms and conditions of access, and use, as indicated by the holding institution and/or rights owner. If a digital database is being created from non-digital materials, it is important to obtain the consent of owners and/or rights’ holders (eg. archives, libraries, publishers) before electronic versions are created.

8(b) Support, software and infrastructure for qualitative data

Support for qualitative data use and elaboration is provided by the EUI Library. Software support is provided by the EUI ICT Service.69 ArcGIS and ATLAS.ti can be used for analysis, mapping and visualisation of qualitative non-numerical data such as audio, graphics, text and video. Coding Analysis Toolkit (CAT) can be used for content and discourse analysis. Tools for data backup (SyncToy), file zipping (7-Zip), data encryption (TrueCrypt) and image adjustment (Resizer) are also available.70 Many of the tools used for the analysis of quantitative data (eg. Gauss, Julia, MATLAB, Python, R, Stata listed in Section 9 below) can also be used to generate aggregate statistical observations from qualitative data.

The analysis of restricted personal data may require scholars to work in a ‘safe-room’ environment. If this is required by a data provider, contact the EUI Library.71

8(c) Research data management and data management plans for qualitative data

Although research data management for qualitative data is similar to research data management for quantitative data – there are some additional considerations. Research data management encompasses the control of data inputs, the handling and use of data, the protection of data, the creation of data outputs and the preservation of data. Research data management is carried out by scholars throughout the duration of a research project (Section 3 above). During data analysis, qualitative data materials

71 econlibrary@eui.eu
should be carefully handled and secured, either in a locked storage unit or in a locked room. This is particularly important for confidential, unique and archival material.

Data management plans (DMPs) are normally required by science funding agencies. Due to the heterogeneous, multi-media and complex nature of qualitative data in the humanities and social sciences, it is particularly important for scholars to keep a record of data sources, including notebooks, questionnaires, codebooks and multilingual thesauri. Supporting documentation serves as the basis for accurate metadata, and facilitates future retrieval and reuse. In the case of non-repeatable, time-sensitive, socio-political research, data management plans may require a detailed explanation of the qualitative research methods used.

8(d) Metadata for qualitative data

Metadata are data about data, presented in a systematic schema. Accurate metadata are necessary for the organisation, use, repositing, sharing and machine-discovery of datasets (Section 5(c) above). The metadata fields used for quantitative data outputs can also be used for qualitative data outputs. However there are additional considerations. As well as the name(s) of scholars and technical collaborators who generate a dataset, it may be necessary to include the authors/creators of subsidiary qualitative data. The dates of creation of subsidiary works included in any new qualitative dataset should be clearly indicated. Linguistic, national and regional metadata should be provided where relevant (e.g. multi-lingual surveys). The format and version of software used to elaborate the data should also be indicated.

8(e) Data preservation, repositing and open qualitative data

Qualitative digital data outputs in the humanities and social sciences can be reposited in the EUI Research Data collection of the Cadmus repository, or in a subject repository or general repository. By carefully noting the metadata elements explained in Section 5(c), scholars will have a ready checklist for determining whether, when, how, where and under what terms, research data outputs can be shared as open data. Qualitative datasets that are made available as open data should be the product of original research and should adhere to best-practice regarding data protection and database copyright (Section 2 above).

9. INFRASTRUCTURE, SOFTWARE AND SUPPORT

The EUI ICT Service provides infrastructure, software and connectivity support. Research software programmes are listed on the ICT web site. Technical support is provided at the site offices of the ICT Service.

The EUI Library maintains the Data Portal, the Micro Data Restricted Server and the EUI Research Data collection in the Cadmus repository; and provides support for data discovery, data access, data use, and helps EUI members prepare data management plans.

In addition to the software programmes for qualitative research, listed at 8(b) above; the EUI ICT Service provides Fortran, Gauss, MATLAB, OxMetrics, Python, R, Stata, Stat/Transfer, WinEdt, WinRATS – and supports the high-performance computing cluster. A guide to the High Performance Cluster (HPC) is available by writing to HPC.support@eui.eu (EUI members only). Advice on the use of statistical software is provided by the ICT Service and software tutors.73

A directory of online research data software manuals, with links to full-text, is available on the Library web site.74 Data software manuals are available in the Badia Library and the Economics Departmental Library (shelfmarks 001 to 006). The Library holds a substantial collection of works on statistical science, data science, applied mathematics, probability, calculus, data mining, machine learning, modeling, optimisation, regression analysis, experimental design, survey methods and time series (shelfmarks 001-006 and 500-519). Books and manuals in any language may be suggested for acquisition by the Library.75

Data support is provided at the Badia Library, and at the Economics Departmental Library, Villa La Fonte (Monday, Wednesday and Friday afternoons from 14:45 to 18:30).

Every Friday during term, the Library issues the Bulletin of Economic Research and Statistical Data. The Bulletin can only be distributed to EUI mail accounts. Sign up by sending a message with 'subscribe' in the title, to econlibrary@eui.eu  Data news is also disseminated via the EUI Library Blog76 and Twitter.77

10. INTERNATIONAL RESEARCH DATA GUIDELINES

Further information is available from these sources:

- Consortium of European Social Science Data Archives (CESSDA)
- Data User Agreements Directory
- Digital Curation Centre: How to develop a data management and sharing plan
- DMPonline data management planner
- EUDAT collaborative data infrastructure
- EUI Research Data
- EUI Research Data reposit instructions
- European Data Portal
- European Open Science Cloud (EOSC)
- GESIS repository - Leibniz Institute for the Social Sciences
- Good Data Protection Practice in Research (EUI)
- Google Dataset Search
- Guide to Social Science Data Preparation and Archiving (ICPSR)
- Horizon 2020 Guidelines on Data Management
- How and Why You Should Manage Your Research Data (JISC)
- MANTRA research data management training
- Metadata for Social Science & Humanities (Digital Curation Centre)

73 https://www.eui.eu/ServicesAndAdmin/ComputingService/Software/ResearchSoftwareTutors
75 Send suggestions to: econlibrary@eui.eu
76 https://blogs.eui.eu/library/category/economics/
77 https://twitter.com/econlibrary
- OECD Guidelines on Research Ethics & New Forms of Data for Social & Economic Research
- Open Data Handbook - Open Knowledge Foundation
- Open Economics Principles - Open Knowledge Foundation
- OpenAire network
- Prepare and Manage Data (UKDS)
- re3data registry of data repositories
- Research Data Alliance (Europe)
- Research Data Curation Bibliography (C.W. Bailey Jr.)
- Research Data Management Toolkit (JISC)
- Social Sciences and Humanities Open Cloud (SSHOC)
- The Hague Declaration on Knowledge Discovery in the Digital Age
- Where to Keep Research Data Checklist (DCC)
- Zenodo data repository (CERN)

Online version of this document:
http://www.eui.eu/Research/Library/ResearchDataServices/Guide.aspx

econlibrary@eui.eu

Thomas Bourke

7th edition

18 March 2019