RESEARCH DATA GUIDE
EUI LIBRARY
9th edition - 2021
This EUI Library Guide provides information and best practices about data discovery, data access, data management planning, data use, data protection, data preservation and data sharing. Information in this guide is intended to help and support all members of the EUI academic community working with data throughout the lifecycle of their research endeavours, from initial data collection, discovery and access up to the sharing of original research data outputs.

1. DATA DISCOVERY AND THE EUI LIBRARY DATA PORTAL

This section describes access to EUI licensed data resources, access to restricted micro-socioeconomic data, access to data hosted by other institutions, and access to open data.

1(a) The EUI Library Data Portal

The Library maintains a Data Portal providing access to licensed resources for EUI members and introduces open data resources of relevance to EUI research themes. All 80 resources indexed in the Data Portal have an online guide with (i) data description and scope (ii) time-period coverage and release/wave information (iii) support links (online manuals, software transfer routines, user networks) and (iv) terms and conditions of access and use. Modes of access for each database are described in the individual resource guides linked from the Data Portal. There are five sub-directories:

- **Macroeconomic, Financial and Historical Data** providing access to national, regional and global economic, political and historical resources.
- **Micro-socioeconomic Data** providing access to household, family, individual and company-level resources.
- **European, EU and Euro Area Data** providing statistics for research on pan-European topics, the European Union, European countries and European sub-state regions.

---

2. [https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/MacroDataSet](https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/MacroDataSet)
- **Data Topics** for matching research themes with data resources\(^5\)
- **Data Resources for Socio-economic Research on COVID-19.**\(^6\)

Databases indexed in the EUI Library Data Portal have one of three modes of access:

- Internet interface via proxy (mostly for macroeconomic, political economic and historical resources)
- Library restricted data server (mostly for micro-socioeconomic resources)
- Desktop application: Datastream via Programmes; Compustat via Stata.

Macroeconomic databases provide statistics on global, regional and national economic and political-economic developments provided by international organisations and major data publishers including Databanks, ECB, Eurostat, Global Financial Data, IMF, OECD, Refinitiv, Standard & Poor’s, World Bank, UN and WIIW.

Micro-socioeconomic databases provide individual, family, household and company observations. Major providers include: DIW, Eurostat, GESIS, ICPSR, Moody’s Analytics-BvD and UKDS.

1(b) Registration for EUI-Library hosted micro-socioeconomic data

EUI members can register for access to micro data hosted by the EUI Library as follows:

- Complete the micro data registration form (selecting the required dataset from the drop-down menu)\(^7\)
- Separate registration is required for each micro dataset
- Registered users are given access to the Library’s restricted data server
- For some providers, parallel registration with data issuers is required
- Specific instructions are provided at 'How to access' on each micro data description page in the Data Portal
- Short-term EUI visitors do not have access to EUI-hosted micro data.

The EUI is a Eurostat-approved research data facility. Access to Eurostat collated data (EU Adult Education Survey, EU Labour Force Survey, EU Statistics on Income and Living Conditions and the European Community Household Panel) requires individual end-user contracts with Eurostat as follows:

- Complete the Library’s Micro Data Registration form (selecting the dataset required from the dropdown menu)\(^8\)
- For the Eurostat contract workflow, create an EU Login\(^9\)
- Note: Eurostat informs that for EUI members - it is necessary to use a non-institutional (non-EUI) personal email account for the creation of the EU Login. (It is not necessary for all members of a research team to create an EU login - only the person submitting the application)

\(^6\) [https://www.eui.eu/Research/Library/ResearchGuides/Economics/Crisis](https://www.eui.eu/Research/Library/ResearchGuides/Economics/Crisis)
\(^7\) [https://www.eui.eu/Research/Library/RequestForms/Register-micro-data](https://www.eui.eu/Research/Library/RequestForms/Register-micro-data)
\(^8\) [https://www.eui.eu/Research/Library/RequestForms/Register-micro-data](https://www.eui.eu/Research/Library/RequestForms/Register-micro-data)
- Project applications are made using the online Microdata Access Workflow Tool\(^\text{10}\)
- When completing the application, EUI researchers should enter the name of their thesis supervisor as Principal Investigator
- Write to econlibrary@eui.eu for (i) the Research Entity ID number and (ii) the description of the EUI’s secure data environment and protocols
- The names and EUI email contact details of all scholars on the project must be included in the form (researchers, fellows, professors)
- The applicant sends the data access application to Eurostat using the Microdata Access Workflow Tool\(^\text{11}\)
- The submitter should monitor their EU Login account for updates regarding the application. (Library staff do not have access to individual EU Login accounts)
- When the application is approved by Eurostat, the applicant should print all documents and ask the Principal Investigator to sign all pages of the application, and sign and date the final page
- The application is sent to the Data Manager for signature\(^\text{12}\)
- Separate confidentiality declarations must be signed by (i) the applicant(s) (ii) the P.I. and (iii) the Data Manager
- The applicant scans the documents and uploads to the EU Login account
- Eurostat confirms receipt. Send this to econlibrary@eui.eu
- The Library provides access to the data via restricted server.

Eurostat informs that no person(s) other than those named in the contract can access Eurostat micro-socioeconomic data. Breach of contract conditions may lead to the withdrawal of the EUI’s research entity recognition and the potential loss of access by other EUI scholars.

1(c) Access to data at other institutions

EUI members who require access to restricted data at another facility (data centre, statistics agency, archive) 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. 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(d) Access to open data resources

Scholars, government agencies and international organisations increasingly share data, codebooks and software via the internet. Open research datasets can be located via Google Dataset Search\(^\text{13}\) and the re3data\(^\text{14}\) registry of research data repositories, indexed by discipline, sub-discipline, data type and host location.

\(^\text{10}\) https://webgate.ec.europa.eu/multisite/microdata/
\(^\text{11}\) https://webgate.ec.europa.eu/multisite/microdata/ Instructions for using the Microdata Access Workflow Tool are at this link: https://ec.europa.eu/eurostat/documents/203647771732/EU_Login_Tutorial/
\(^\text{12}\) econlibrary@eui.eu
\(^\text{13}\) https://datasetsearch.research.google.com/
\(^\text{14}\) http://www.re3data.org/
2. DATA PROTECTION, DATABASE COPYRIGHT AND ETHICAL USE

This section provides an overview of data protection, the processing of sensitive data, EUI Ethics Committee review, security protocols, terms and conditions of use, database copyright and individual data user undertakings.

2(a) Data protection

The EUI ‘Guide to Good Data Protection Practice in Research’ provides details of data protection at the EUI. Special terms and conditions apply to access and use of micro-socioeconomic and qualitative data. This reflects the sensitive nature of observations about human subjects, families and households. Such terms and conditions apply to (i) data collected by users during research projects (ii) data provided to EUI members by the Library (iii) data provided directly to EUI members under third-party license. Terms and conditions of access and use for all datasets hosted by the Library are provided in the ‘full details’ sections of the resource guides linked from the Data Portal. EUI members accessing data directly from a third-party provider should read and comply with the stipulated terms and conditions of access and use.

Persons, families and households cannot be identifiable in any dataset. When using micro-socioeconomic datasets, users must not attempt to identify any individual, family or household. The collection, processing, use and storage of personal data is subject to data protection rules. The processing of sensitive data relating to health, racial/ethnic origin, political opinions, religious beliefs, sexual orientation, genetic/biometric observations, memberships &c. requires even stricter safeguards. The EUI ‘Guide to Good Data Protection Practice in Research’ states: “As a rule, the processing of sensitive data is prohibited. However, Article 8 of the EUI’s Data Protection Policy provides for specific circumstances which allow for the processing of sensitive data. The most common in research is upon the data subject’s explicit consent.”

When collecting, processing, using and storing personal data; scholars are responsible for obtaining and documenting the informed consent of subjects. Consent must be intelligible; it should refer clearly and precisely to the scope and the consequences of the data processing and the retention period. EUI scholars can use the sample consent form. Where written consent is not possible, audio or video recording may be used.

Throughout the research data cycle, scholars are responsible for preserving the confidentiality of data observations pertaining to human subjects, families and

---

households. If necessary, data should be encrypted (eg. by using TrueCrypt). The use of cloud computing services must not reduce the level of data protection.

2(b) Anonymisation and Pseudonymisation

Data which is presented for sharing via a data repository must be anonymised so that persons, families and households cannot be directly or indirectly identified. Anonymisation techniques include: the removal of direct identifiers from datasets (names, postcodes, workplaces, telephone numbers &c.); the aggregation, banding or reduction of the precision of variables (eg: year of birth rather than year/month/date of birth); the reporting of general levels of observation (eg. general geo-codes rather than specific geo-codes); the generalisation of categories (eg. sector of employment rather than precise employment function); and the removal of the upper or lower ranges of observations.

Pseudonomysiation is a technique used when it may be necessary to retain personal data for future processing (eg. information required for a new survey wave). Pseudonomysiation entails the substitution of personally identifiable information with a unique identifier not connected to the human subject’s real identity - using techniques such as coding. The data and the key file must be kept separate and secure, and must be destroyed at the end of the retention period of the research project.

2(c) Processing of sensitive data, EUI Ethics Committee review and security protocols

The collection and processing of personal data for research purposes may require EUI scholars to apply for Ethics Committee approval before the data phase of the project can begin. Principal Investigators should consult the initial checklist and supporting documentation on the EUI Ethics and Integrity in Academic Research web page.21 Ethics Committee reviews should be initiated before data collection begins, because approval cannot be granted retroactively. The Ethics Committee review can take up to six weeks.

When data suppliers stipulate extra security measures for data access and elaboration, EUI members should consult the Library’s support page for the Use of Sensitive Data.22 The following information should be sent to econlibrary@eui.eu: (i) name of the data issuer (ii) title of the dataset with URL (if available) and (iii) names of all project participants. The following questions should be addressed: Does the data issuer require that the data be accessed and used on a stand-alone off-line computer? Does the data issuer require that the data be used in a secured room? Does the data issuer require a contract with the individual end-user(s)? Does the data issuer require an institutional EUI contract (‘guarantor’) and/or signature of the EUI Data Protection Officer? Are there any special conditions regarding access to the data (eg. remote protocol)? Will any external (non-EUI) project collaborators be working with the data? If required, the ICT Data Security Officer will provide a description of ICT protocols and infrastructure.

2(d) Terms and conditions of use and database copyright

Access to, and use of, databases provided by the EUI Library are subject to contractual

---

22 https://www.eui.eu/Research/Library/ResearchDataServices/SensitiveData
license agreements and database copyright terms. Full details are on the Library’s Terms and Conditions’ web page.\textsuperscript{23} 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 and may expose the university to financial penalties.

EUI 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 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 required by the license.\textsuperscript{24}

\textbf{2(e) Agreements between individual EUI members and third-party data providers}

Some data issuers require that access contracts be established directly with end-users. Such contracts may require the counter-signature of an EUI administrator. If a data issuer requests a counter-signature, write to the Library.\textsuperscript{25}

\textbf{3. DATA MANAGEMENT PLANS (DMPs)}

Data management plans are short documents, normally required by research funding agencies when applying for research grants. EUI members who are required to submit a data management plan - either as part of a funding application, or during a research project - should contact the EUI Library for assistance. The online tools described in Section 3(b) can be used to generate a DMP.

\textbf{3(a) Generating a data management plan}

Data management plans are short documents which describe:

- How data is generated or collected
- 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 dataset authorship and credit are assigned
- How data is preserved
- How, whether and under what terms research data outputs can be shared.

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 might be shared under more restrictive terms and conditions. The FAIR data principles - to make data findable, accessible, interoperable and reusable - must be considered.

\textsuperscript{23} \url{http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx}
\textsuperscript{24} \url{http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx}
\textsuperscript{25} econlibrary@eui.eu
during the preparation and revision of data management plans.\footnote{https://www.go-fair.org/fair-principles/}

If the Principal Investigator is not the same person as the project Data Manager, this should be stated in the DMP. In international collaborative projects, the name of the person who has final authority regarding decisions on credit for dataset creation, and the sharing of data outputs, should be indicated. Unless otherwise stated the Principal Investigator undertakes this role. Science funders normally require an initial DMP, 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 to incorporate new data inputs, generation and methodology, and relevant changes to the composition of the research team or consortium.

3(b) Data management planning tools

Online tools can be used to prepare structured data management plans - complying with EU Horizon Europe (2021-2027), European Research Council and other science funder DMP requirements.

- **DMPonline\footnote{http://www.dcc.ac.uk/dmponline}**\footnote{https://www.dcc.ac.uk/dmponline} (maintained by the Digital Curation Centre) can be activated via 'Create Account' at the upper right of the homepage. For organisation enter 'other’. The Principal Investigator (P.I.) should be identified in the data management plan. For research teams, the P.I. can assign co-author plan editing rights by entering colleagues' email addresses and assigning the status of 'co-owner’, ‘editor’ or ‘read only.’ Enter project details (title, abstract, start/end dates &c.) and click on ‘Initial DMP’ in the top menu. Complete the sub-fields to generate the data management plan.

- **Argos\footnote{https://argos.openaire.eu/splash/}**\footnote{https://argos.openaire.eu/splash/} is an online DMP creation tool maintained by OpenAIRE. Log onto the Argos platform with an email address and click 'Launch Wizard' and follow the step-by-step instructions.

- **The DMP Evaluation Rubric\footnote{https://scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf}**\footnote{https://scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf} developed by Science Europe in 2021, provides core criteria which can be used by principal investigators and project managers to develop and evaluate data management plans.


- **FAIR-Aware\footnote{https://fairaware.dans.knaw.nl/}**\footnote{https://fairaware.dans.knaw.nl/} maintained by the Data Archiving and Networked Services Institute is an online checklist tool to help researchers build compliance with the FAIR data principles.


3(c) Support for EUI principal investigators and project managers

EUI members can use the following prompts when compiling a data management plan. EUI members can send their draft DMPs to the Library for input, feedback and review.

<table>
<thead>
<tr>
<th>Overview of research project</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Summary of the research project, including discipline/sub-discipline, scope and methodology</td>
</tr>
<tr>
<td>• Overview of data collection and generation</td>
</tr>
<tr>
<td>• Type of data being collected, generated and used</td>
</tr>
<tr>
<td>• Origin of data inputs.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anticipated data outputs and utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>• List of expected data outputs</td>
</tr>
<tr>
<td>• Expected community utility; eg. academic researchers, policymakers, media, general public</td>
</tr>
<tr>
<td>• Possible inter-disciplinary use</td>
</tr>
<tr>
<td>• Data quality assurance processes.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resources required for data collection, generation, use and preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Expected financial costs associated with accessing and generating data, where applicable</td>
</tr>
<tr>
<td>• Anticipated data support, technical and research assistance requirements</td>
</tr>
<tr>
<td>• Data management responsibilities of research project leaders, team members and partners.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data security, infrastructure and protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>• EUI members should write to <a href="mailto:econlibrary@eui.eu">econlibrary@eui.eu</a> to request the Library/ICT standard description of data security, infrastructure and protocols. This internal document describes layered physical security, network security, authentication protocols and other elements for inclusion in a data management plan.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethical considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Pre-project Ethics Committee review may be required for survey, ethnographic, qualitative, experimental and other data collection, generation and processing</td>
</tr>
<tr>
<td>• Description of data protection measures guaranteeing that persons, families and households are not identifiable in dataset outputs</td>
</tr>
<tr>
<td>• List pre-existing data resources to be used, and acknowledge database copyright compliance.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FAIR Data Principles making data</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Findable (ii) Accessible (iii) Interoperable and (iv) Reusable</td>
</tr>
</tbody>
</table>
Making data Findable
- Datasets should be assigned accurate and consistent metadata to aid findability and machine retrievability
- Folders, files, variables and versions should be consistently named to aid discovery
- Datasets should have persistent, unique digital object identifiers (eg. DOIs)
- EUI members can refer to Section 6(c) of this guide for metadata schema details.

Making data Accessible
- Status ‘open’ or ‘restricted’ should be assigned to research dataset outputs
- If access to data outputs will be restricted, reasons should be given
- If access to restricted data can be requested, information about the procedure should be provided
- The repository where the data will be made available should be identified (eg. Cadmus, the EUI research repository)  
  33
- Software packages that can be used to access the data should be listed.

Making data Interoperable
- Data vocabularies, variable schemas and other standards should be explicit, and be readable by standard software packages to facilitate data reuse
- Standard metadata and naming conventions should be used
- The Dublin Core schema - used for the Research Data Collection in Cadmus, the EUI research repository - facilitates interoperability
- Cadmus, the EUI research repository, is built using web standards and provides both human- and machine-readable interfaces to search, discover and access the reposited data.

Making data Reusable
- Open data generated by the project should be made available under an open licence that supports, and clearly states, the reuse conditions - for example the Creative Commons Attribution International (CC-BY International) or Creative Commons Public Domain Universal (CC0) licenses
- Supporting documentation and codebooks should be reposited with the data, to make the data comprehensible and useful for other stakeholders
- Information about any tools or instruments necessary to reuse or verify the data should be provided
- If there will be a data embargo period after the research project, this should be noted.

EUI research data guidelines and standards

- EUI members who are required to refer to institutional guidelines can cite these documents:
  - EUI Library ‘Research Data Guide’ (9th ed) 2021
  - EUI ‘Code of Ethics in Academic Research’ 2021

33 https://cadmus.eui.eu/
34 https://creativecommons.org/licenses/by/4.0/
35 https://creativecommons.org/publicdomain/zero/1.0/
36 https://www.eui.eu/Research/Library/ResearchDataServices/Guide
4. RESEARCH DATA MANAGEMENT IN HORIZON EUROPE AND OTHER EU PROGRAMMES

EUI project managers and principal investigators who are preparing EU Horizon Europe (2021-2027) and European Research Council (ERC) grant proposals are required to submit preliminary information about data management provisions in the proposed project. The Draft Model Grant Agreement for EU Horizon Europe funded research projects was published in February 2021, with an overview of research data management provisions on pp.108-109. The Horizon Europe Programme Guide – due for publication in June 2021 – will provide further details of data management obligations for Horizon Europe project managers.

Data management sections of project proposals are evaluated by the European Commission under the criterion ‘impact’, which includes scientific and societal impact. When completing the general information section of EU funding applications, principal investigators and project managers should address: (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 outputs cannot be made available, explain why (v) How will the data be curated and preserved? The EUI Library assists project managers and principal investigators with both the data management section in funding proposals and the subsequent data management plans.

When an EU project proposal is approved, a preliminary data management plan (DMP) must be provided within six months of the start of contract. The European Commission mandates two further versions of the DMP: one at project mid-point and one as a final deliverable. These further versions are intended to address the evolving nature of the research project, with the final one being considered as a formal commitment. The DMPonline and Argos data management planning tools (Section 3(b) above) can be used to generate EU compliant data management plans.

In the Horizon Europe programme, open science – which includes open access to research data – is evaluated under the ‘excellence’ criterion. The European Commission’s recommendation for sharing research data outputs is: “As open as possible; as closed as necessary”. The Horizon Europe programme (2021-2027) requires grant beneficiaries to deposit research data outputs in a research data repository which is federated with the European Open Science Cloud (EOSC, see 4(a) below). Cadmus, the EUI research repository, can be used for the reposit of EUI Horizon Europe research datasets. The EC 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.”

Project results can also be shared via the Horizon Results

---

39 Further information is also contained in these Horizon Europe documents: Model Grant Agreement (MGA) article 17; Work Programme General Annexes; Proposal template and; Annotated Grant Agreement (AGA), article 17.
40 econlibrary@eui.eu
42 https://cadmus.eui.eu/
Platform, launched in April 2021.44

4(a) European Open Science Cloud (EOSC)

The European Open Science Cloud portal was launched in 2018.45 The EOSC portal enables the discovery of data, provides tools for data analysis and information about storage, computation, training and security. Content and services are available via the EOSC catalogue.46 The European Commission has published two supporting documents: Prompting EOSC in Practice47 which explains the governance of the new service, and Turning FAIR into Reality48 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.

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

The Social Sciences and Humanities Open Cloud initiative was launched in 2019. SSHOC constitutes the social sciences and humanities cluster of the European Open Science Cloud and “aims to bring together existing and new infrastructures from social science and humanities ERICs (European Research Infrastructure Consortiums) and foster interdisciplinary research and collaboration.”49 There are 47 institutional participants, under the coordination of the Consortium of European Social Science Data Archives (CESSDA).

5. 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, security and backup during the project, folder structure, file structure, file naming, variable naming, software format, documentation and codebooks, data protection, informed consent of data subjects, anonymisation, copyright, preservation, archiving, data citation and - where possible - sharing data outputs as open data. The Library provides assistance with all of the above. The following guides are useful for scholars working in the social sciences and humanities:

- CESSDA Data Management Expert Guide50
- JISC Research Data Management Toolkit51
- UKDS Prepare and Manage Data Guide.52

---

44 https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/horizon-results-platform
45 https://eosc-portal.eu/
46 https://eosc-portal.eu/
49 https://www.cessda.eu/About/Projects/Current-projects/SSHOC
51 https://rdmtoolkit.jisc.ac.uk/
52 https://www.ukdataservice.ac.uk/manage-data/
5(a) Data input and quality control

The collection and generation of data should be informed by high standards of quality control. During the research project, it is important to keep an accurate and updated record of data sources and data generation. Datasets generated from empirical work during a research project should adhere to the ethical best-practices of the discipline or sub-discipline, with close attention to accuracy, completeness and consistency. Data documentation should be updated throughout the research project. Scholars undertaking surveys and experiments should ensure that anonymisation techniques and pre-agreed terms of disclosure are used and documented (Section 2(b)). 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.

5(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 make research data findable, accessible, interoperable and re-usable (FAIR principles).53

The folder structure of research datasets (eg. hierarchical/horizontal) should be considered early in the project. Qualitative datasets containing text, interviews, images &c. may require individual files for every element. File names should be standardised and consistent, eg: 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 (eg. YYYY-MM-DD) or other consistent versioning. Changes to the structure should be documented.

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 open source accessibility. Details of how to submit datasets to the Research Data collection in Cadmus, the EUI research repository are in Section 6.

5(c) Documentation and codebooks

Clear and accurate documentation should be provided about the purpose, context and methodology of the research project and research data outputs. Good documentation makes datasets findable, accessible, interoperable and re-usable (FAIR principles). Documentation should include a detailed description of the dataset, providing information about folders, files, variables, versioning and - where applicable - information about problematic values, missing observations and weightings. Codebooks, questionnaires and data dictionaries should be included.54 A concise note on methodology should be

54 https://www.icpsr.umich.edu/web/pages/deposit/guide/
provided, along with information on how and when the data was collected or generated, and how the data was elaborated during the research project.

5(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. Preliminary findings and associated documentation relating to sensitive data 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.55

6. REPOSITING AND PRESERVING EUI RESEARCH DATA IN THE CADMUS REPOSITORY

The EUI Research Data Collection within Cadmus, the EUI research repository, is an infrastructure for the preservation of EUI research data outputs, a platform for sharing research data, and a solution for funder-mandated research data management compliance (eg. the EU Horizon Europe and European Research Council programmes).

6(a) Preparing data for reposit in Cadmus, the EUI research repository

When preparing to submit a research dataset for inclusion in Cadmus, it is advisable to consider and prepare for the submission of both data file(s) and documentation file(s). Large datasets should be submitted in a compressed archival format (eg. .zip, .tgz). Data should be submitted in the original file format version. Subsets must be accommodated within a folder structure within the archival file – not submitted 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 and submitted along with the data files. Scholars presenting datasets for inclusion in the Research Data Collection in Cadmus should pay particular attention to data quality control, dataset structure and data protection.

6(b) Completing the dataset submission form

All EUI researchers, fellows or professors from any discipline can submit research data outputs to the Research Data Collection in Cadmus, the EUI research repository. The first step is to complete the online submission form.56 Library staff will use the submitted information to generate a formal metadata description in the Dublin Core schema that will be imported into Cadmus and will generate the appropriate description page for the dataset. This process is essential to render the data findable, accessible, interoperable and reusable (FAIR data principles). Library staff will then make an appointment for data transfer.

Datasets presented for inclusion in the Cadmus repository must be the output of research by a current EUI member, or an EUI research team, or a team of researchers

55 https://www.eui.eu/ServicesAndAdmin/ComputingService/PolicyDocuments/StrongPasswordPolicy
56 https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
with at least one EUI member. The name of the Principal Investigator, researcher(s), and 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 submitting the online data submission form, the EUI member acknowledges that the dataset presented for inclusion in the EUI Research Data Collection is the output of original data generation and/or collection; or the output of significant, value-added, elaboration of pre-existing sources. If the dataset is the output of original data collection and elaboration, details must be provided. If the dataset is derived from pre-existing sources, these must be indicated (data creator, institutional source, publisher). 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 also assists scholars to reposi datasets in discipline-specific data repositories. Major data repositories are indexed in the international re3data registry and Google Dataset Search.

EUI members submitting data to the Cadmus repository should state whether or not the dataset can be shared as open data or will 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 limited public version of a larger restricted dataset. Embargo status can change over time.

6(c) Metadata

Metadata are ‘data about data’ presented in a formal 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.

Approaching the end of a research project, 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. Dataset metadata in the EUI Research Data Collection are generated by Library staff, using information from the online submission form. These are the principal 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.).

57 http://www.re3data.org/
58 https://datasetsearch.research.google.com/
59 http://www.dcc.ac.uk/resources/subject-areas/social-science-humanities
60 https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
Where researcher IDs are available, eg. ORCID, these should be provided.

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

**TITLE OF DATASET**
The title should succinctly convey the nature and scope of the dataset. (This should not be identical to the project title, or a related publication title.)

**DESCRIPTION OF DATA**
A meaningful abstract, describing the data, the purpose and scope of the research project, and the methodology, must be provided. The Library can help edit data 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 pre-existing database(s)*, all source(s) must be clearly cited.

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

**YEAR OF COMPLETION OF DATASET**
The date of completion of the dataset must be provided. If the dataset is part of a series, 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 file format and version must be given (eg. .csv, Excel, Stata 17, .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.

**LICENSE**
Dataset submitters should choose a license option from the drop-down menu: CC-BY (international)\(^61\) or CC0.\(^62\)

\(^61\) [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)
\(^62\) [https://creativecommons.org/share-your-work/public-domain/cc0/](https://creativecommons.org/share-your-work/public-domain/cc0/)
DIGITAL IDENTIFIER
Library staff will assign a unique object identifier to the dataset for the purpose of discovery, linking and citation.

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. (The Library can also cross-index datasets and publications at a later date.)

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

DATASET CITATION
The Cadmus repository will generate a standardised citation for the dataset, eg:

When submitting details of research publications to the EUI Cadmus repository, EUI scholars should cite datasets which support the research findings. Further information about data citation is provided by DataCite.

7. OPEN DATA

Scholars, government agencies and international organisations increasingly share datasets, codebooks and software via the internet. Major data repositories are indexed in the international re3data registry and Google Dataset Search.

By carefully generating the metadata elements listed in Section 6(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 at the conclusion of a research project. 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(d)). Library staff can help EUI members determine whether and when a dataset can be openly shared.

63 https://cadmus.eui.eu/
64 DataCite metadata schema: https://schema.datacite.org/
65 http://www.re3data.org/
66 https://datasetsearch.research.google.com/
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 reposited 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 (eg. history, anthropology, ethnography) - especially where there is a mix of qualitative and quantitative methods. It is important that dataset design is located in the culture of the discipline or sub-discipline in which the research is undertaken.

8(a) Terms and conditions of use for qualitative data

Qualitative data can be generated from surveys, interviews, free-text responses to questionnaires, focus group data, ethnographic observation and/or recordings and experimental simulations. Human subjects should be informed of their rights as established by jurisdictional data protection legislation and the best practices established by scholarly societies in the relevant discipline. Scholars should consider risks to personal data subjects before embarking on personal data gathering (eg. via interviews or questionnaires). When collecting, storing, using or transferring personal data, scholars are responsible for obtaining the informed consent of subjects.67 As with the use of micro-socioeconomic data; scholars using qualitative data should pay particular attention to ethical standards when handling data pertaining to health, racial/ethnic origin, political opinions, religious beliefs, sexual orientation, genetic/biometric observations, memberships &c. Persons, families and households cannot be identifiable in any dataset. When creating a digital database from non-digital materials (eg. photographs of primary materials), it is important to obtain the consent of rights’ holders and/or hosting facilities (archives, libraries, museums &c.).

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.68 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.69 Many of the tools used for the analysis of quantitative data (eg. Gauss, Julia, MATLAB, Python, R, Stata listed in Section 9) 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. 70

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. During data analysis, qualitative data materials (eg. primary sources) should be carefully handled and secured. This is particularly important for confidential, unique and archival material which should be stored in a properly secured space, either digital or physical.

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 require a detailed explanation of the qualitative research methods used.

8(d) Metadata for qualitative data

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 (eg. multi-lingual surveys). The format and version of software used to elaborate the data should also be indicated. The Oral History Association (OHA) maintains an online Manual of Best Practices for Archiving Oral History, adopted in October 2019, providing good practices for metadata and description. 71

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

Qualitative digital data outputs in the humanities and social sciences can be reposited in the Research Data Collection in Cadmus, the EUI research repository, or in a subject repository, or in a general purpose repository such as Zenodo. 72 By carefully noting the metadata elements explained in Section 6(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. Some kinds of ethnographic and other qualitative data may require special archival access controls and protocols.

70 https://www.eui.eu/Research/Library/ResearchDataServices/SensitiveData
72 https://zenodo.org/
9. EUI 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 Data Server and the Research Data Collection in Cadmus, the EUI research repository. The Library also 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 in Section 8(b); 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.

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

Data support is provided at the Badia Library (BF-085) and at the Economics Departmental Library, Villa La Fonte (VLF-026).

Every Friday during term, the Library issues the ‘Bulletin of Economic Research and Statistical Data.’ EUI members can sign up by sending a message with ‘subscribe’ in the title to econlibrary@eui.eu The Bulletin can only be distributed to EUI mail accounts. Data updates are also disseminated via the EUI Library Blog and Twitter. Other useful Library resources include the Publishing and Open Sciences web page – especially the section Manage Your Research Data.

---

75 [https://www.eui.eu/Research/Library/ResearchGuides/Economics/Software](https://www.eui.eu/Research/Library/ResearchGuides/Economics/Software)
76 econibrary@eui.eu
77 [https://blogs.eui.eu/library/category/economics/](https://blogs.eui.eu/library/category/economics/)
78 [https://twitter.com/econlibrary](https://twitter.com/econlibrary)
79 [https://www.eui.eu/Research/Library/EUIPublications/RDM](https://www.eui.eu/Research/Library/EUIPublications/RDM)
80 [https://www.eui.eu/Research/Library/EUIPublications](https://www.eui.eu/Research/Library/EUIPublications)
10. INTERNATIONAL RESEARCH DATA GUIDELINES

This guide draws on many diverse international sources – in particular those listed below, which are strongly recommended to scholars working with research data.

- Argos online DMP creation tool (OpenAIRE)
- CESSDA Data Management Expert Guide
- DMPonline data management planner
- DMP Evaluation Rubric (Science Europe)
- EUDAT collaborative data infrastructure
- EUI Library Data Portal
- EUI Research Data Collection in Cadmus
- EUI Research Data submission form
- European Data Portal
- European Open Science Cloud (EOSC)
- F-UJI automated FAIR data assessment tool
- 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 Europe Guidelines on Data Management
- How and Why You Should Manage Your Research Data (JISC)
- LIBER data management plan catalogue
- MANTRA research data management training
- Metadata for Social Science & Humanities (Digital Curation Centre)
- OECD Guidelines on Research Ethics & New Forms of Data
- Open Data Handbook - Open Knowledge Foundation
- OpenAire network
- Prepare and Manage Data (UKDS)
- re3data registry of data repositories
- Research Data Alliance
- 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)