European University Institute

LIBRARY

# Research Data Guide EUI Library

## 8th edition
## 2020

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 in the Cadmus repository
6. Research Data in EU funded projects
7. Open Data
8. Qualitative data in the humanities and social sciences
9. EUI infrastructure, software and support
10. International research data guidelines

This EUI Library Guide provides information about data discovery and access; 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 requiring ethics review; access to data hosted by other institutions; and access to open data resources.

### 1(a) The EUI Library Data Portal

The Library maintains a Data Portal[1] which provides access to licensed resources for EUI members and introduces open data resources of relevance to EUI research themes. All 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[2]
- Micro-socioeconomic Data providing access to household, family, individual and company-level resources[3]
- European, EU and Euro Area Data providing statistics for research on pan-European topics, EU states and European sub-state regions[4]
- Data Topics Directory helping match research topics with data resources[5]
- Crisis Data Directory presenting major resources for research on socio-economic impacts of the COVID-19 pandemic.[6]

The Crisis Data Directory has three sections: (i) Socioeconomic dimensions: households, families, labour markets, work, unemployment, health, consumption and domestic

[1] https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal
[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
[5] https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/Topics
[6] https://www.eui.eu/Research/Library/ResearchGuides/Economics/Crisis

finances (ii) Macroeconomic, financial system and Euro area stability: markets, currencies, bonds, debt, sectoral transformation and recovery and (iii) Data for comparative crisis research: long-range time-series, historical national accounts, GDP, exchange rates, inflation and economic cycles.

1(b) Access to Sensitive Data: ethics review, DPO oversight and isolated computers

Some data providers require EUI Ethics Committee approval and/or Data Protection Officer clearance. Principal Investigators should initiate the request for approval, providing (i) the name of the data issuer (ii) title of the dataset required with URL (if available) and (iii) names of all project participants. The following information should be sent to econlibrary@eui.eu: 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 DPO? Are there special conditions regarding access to the data (eg. remote protocol)? Will any external project collaborators be working with the data? If required by the data issuer, the ICT Data Security Officer will provide a description of ICT protocols and infrastructure.[7]

1(c) Access to EU micro-socioeconomic data provided by Eurostat

In January 2020, Eurostat revised the application process for access to micro-socioeconomic data. EUI members can apply for access under the Library's Eurostat approved facility contract: (i) Complete the Micro-socioeconomic Data Registration Form[8] (selecting the required dataset from the dropdown menu) (ii) The EUI researcher - or Principal Investigator (or delegate in a research team) - creates an online EU Login[9] (iii) Eurostat informs that for EUI researchers - it is necessary to use a personal (non-institutional, non-EUI) personal email account for the creation of the EU Login (only the person submitting the application needs an EU Login account) (iv) Project applications are made using the online Microdata Access Workflow Tool[10] (v) Before completing the application, consult the individual Eurostat dataset pages on the Library Data Portal for details of variables, reference years and population which are required for the online form (vi) The 'Principal Researcher' is normally an EUI Professor or Post-doctoral fellow (vii) For 'Data Manager' enter Thomas Bourke (viii) Write to econlibrary@eui.eu requesting the EUI Research Entity ID (ix) The names and EUI email contact details of all researchers associated with the project must be included (x) The applicant sends the data access application to Eurostat via the Microdata Access Workflow Tool (xi) The submitter should monitor the EU Login account for updates regarding the application (xii) After approval by Eurostat, the applicant prints the contract; the P.I. initials all pages of the application and signs, dates the final page; the Data Manager (econlibrary@eui.eu) signs the application; separate confidentiality declarations must be signed by the applicant(s), the P.I. and the Data Manager (xiii) The applicant must scan the documents and upload to the EU Login account (xiv) The Library provides access to the data via

[7] https://www.eui.eu/Research/Library/ResearchDataServices/SensitiveData
[8] https://www.eui.eu/Research/Library/RequestForms/Register-micro-data
[9] https://webgate.ec.europa.eu/cas/about.html
[10] https://webgate.ec.europa.eu/multisite/microdata/content/workflow-system-handling-microdata_en

restricted server.

1(d) 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.[11] 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(e) 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[12] and the re3data[13] registry of research data repositories, which is indexed by discipline, sub-discipline, data type and host location. 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.

2(a) Data protection

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 gathered by users during research projects, (ii) data provided 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 all datasets hosted by the Library, are provided in the 'full details' sections of the resource guides linked from the Data Portal.[14]

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 storage, transfer and use of personal data is subject to data protection rules.[15] The processing of particularly sensitive data (eg. pertaining to ethnicity, health, religion, political or sexual orientation, biometrics &c.) requires even stricter

[11] econlibrary@eui.eu (regarding Data Portal access) resdata@eui.eu (regarding the reposit of data in Cadmus)
[12] https://datasetsearch.research.google.com/
[13] http://www.re3data.org/
[14] https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal
[15] For the EUI, see the President's decision on Privacy: https://www.eui.eu/Documents/AboutEUI/Organization/PresidentDecisionOnPrivacy.pdf and the Guide on Good Data Protection Practice in Research): https://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/ResearchEthics/Guide-Data-Protection-Research.pdf

safeguards.[16] 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."[17]

When collecting (eg. via surveys, 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. Consent must be documented by the researcher.

EUI members should use the online Micro Data Request form[18] when applying for access to micro-socioeconomic data hosted on the Library's restricted data server. A separate form is required for each 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 at Villa La Fonte (VLF, 035). Some providers require the submission of a detailed project proposal, and parallel registration, before access is granted. Non-EUI members and short-term visitors do not have access to EUI-hosted micro data.

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 database copyright terms. Full details are on the Library's Terms and Conditions' web page.[19] Data users are individually responsible for compliance with terms and conditions of access and use. 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. Violation of license terms puts at risk other EUI members' future access to data resources.

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 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.[20]

[16] EUI Guide to Good Data Protection Practice in Research, p.7: http://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/ResearchEthics/Guide-Data-Protection-Research.pdf
[17] EUI Guide to Good Data Protection Practice in Research, p.9: https://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/ResearchEthics/Guide-Data-Protection-Research.pdf
[18] https://www.eui.eu/Research/Library/RequestForms/Register-micro-data
[19] http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx
[20] http://www.eui.eu/Research/Library/ElectronicResources/TermsAndConditions.aspx

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 requests the counter-signature of an EUI administrator, please write to the Library.[21]

## 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 the research project lifecycle, it is important to keep an accurate record of data inputs, dataset design, folders, files, variables and versioning. Scholars undertaking experiments should ensure that anonymisation techniques and pre-agreed terms of disclosure are documented and preserved. The data and methodology should be clearly documented. Where applicable, code should be preserved for reposit and sharing. Metadata - 'data about data' - should be assigned precisely (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 should be shared under more restrictive terms and conditions. The FAIR data principles - designed to make data findable, accessible, interoperable and re-usable - must be considered during the preparation and revision of data management plans.

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 DMPonline[23] - the Digital Curation Centre's data management planning tool - can be used to prepare a structured data management plan, complying with EU Horizon 2020 and EU Horizon Europe (2021-2027) 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

[21] econlibrary@eui.eu
[22] resdata@eui.eu
[23] http://www.dcc.ac.uk/dmponline

assign co-author rights, by entering colleagues' email addresses and assigning the status of '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; eg. indicating new data inputs and generation or changes to 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; anonymisation; copyright; preservation; archiving; data citation and - where possible - sharing data outputs as open data.

The following guides are useful for scholars working in the social sciences and humanities: the CESSDA Data Management Expert Guide[24] the JISC Research Data Management Toolkit[25] and the UKDS Prepare and Manage Data Guide.[26]

### 4(a) Data input and quality control

The generation and/or sourcing of data should be informed by high standards of quality control. Datasets 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. Changes should be documented throughout the research project.

### 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

[24] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide
[25] https://rdmtoolkit.jisc.ac.uk/
[26] https://www.ukdataservice.ac.uk/manage-data/

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).[27]

The folder structure of the research dataset should be considered at the beginning of the project (eg. hierarchical / horizontal). 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 (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 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 the research data output. 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.[28] A concise note on methodology should be provided, along with information on how and when the data was generated or collected, and how the data was collated and elaborated during the research project.

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. 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.[29]

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, generation and use* throughout the research cycle. Further

[27] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[28] https://www.icpsr.umich.edu/web/pages/deposit/guide/
[29] https://www.eui.eu/ServicesAndAdmin/ComputingService/PolicyDocuments/StrongPasswordPolicy

information is provided in the EUI Guide to Good Data Protection Practice in Research.[30] 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 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); and banding.[31] 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 database copyright.[32]

## 5. REPOSITING AND PRESERVING RESEARCH DATA IN THE CADMUS REPOSITORY

EUI members can preserve and share their research data outputs by submitting them for reposit in the EUI Research Data collection of the Cadmus repository.[33] This section of the Guide provides information on preparing data for long-term preservation and, where possible, sharing outputs as open data. Section 5(c) below, explains how to create metadata – which are essential for dataset reposit, retrieval and reuse.

### 5(a) Preparing data for reposit

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.[34] Library staff will use the information from the form to generate metadata in the Dublin Core schema.[35] 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 a 'Docs' folder in the repository.

---

[30] https://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/ResearchEthics/Guide-Data-Protection-Research.pdf
[31] Anonymisation techniques are described in Eurostat's Ethics and Data Protection guidelines (Section #3, pp.7-8) https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf, the UKDS Guidelines to Anonymisation for quantitative and qualitative data: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation and the EUI Guide on Good Data Protection (p.16), http://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/ResearchEthics/Guide-Data-Protection-Research.pdf
[32] econlibrary@eui.eu
[33] https://cadmus.eui.eu/handle/1814/64544
[34] https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
[35] http://dublincore.org/metadata-basics/

It is not possible to reposit unstructured data elements and observations in the repository. 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.

5(b) Completing the dataset submission form

By submitting the online form[36] EUI members acknowledge that the dataset for reposit 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 an EUI research team, or a team of researchers with at least one EUI member. 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.[37] 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 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.

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.[38] Major data repositories are indexed in the international re3data registry[39] and Google Dataset Search.[40]

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

[36] https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
[37] http://www.eui.eu/Documents/ServicesAdmin/DeanOfStudies/CodeofEthicsinAcademicResearch.pdf
[38] http://zenodo.org/about
[39] http://www.re3data.org/
[40] https://datasetsearch.research.google.com/

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.[41]

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 are generated by Library staff using information from the online submission form.[42] 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.). Where researcher IDs are available, eg. ORCID, these should be provided. The CRediT framework provides guidelines on authorship and credit for research outputs.[43]

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 a pre-existing database*, 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.

[41] http://www.dcc.ac.uk/resources/subject-areas/social-science-humanities
[42] https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow
[43] http://casrai.org/credit

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. .csv, Excel, Stata 16, .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 chose a license option from the drop-down menu: CC-BY (international)[44] or CC0.[45]

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: 'REILJAN, Andres, FERREIRA DA SILVA, Frederico, CICCHI, Lorenzo, GARZIA, Diego, TRECHSEL, Alexander H., EU Profiler/euandi trend file (2009-2019), EUI Research Data, 2020, Robert Schuman Centre for Advanced Studies.' This can be used to cite datasets in EUI theses, working papers, articles, books and other publications. When submitting details of research publications to the EUI Cadmus repository, EUI scholars should cite datasets which support the research findings.[46] Further information about

[44] https://creativecommons.org/licenses/by/4.0/
[45] https://creativecommons.org/share-your-work/public-domain/cc0/
[46] http://cadmus.eui.eu/

data citation is provided by DataCite.[47]

## 6. RESEARCH DATA IN EU FUNDED PROJECTS

EUI project managers who are preparing EU Horizon 2020 and Horizon Europe (2021-2027) applications are required to submit preliminary information about data management in the proposed project.[48] (Detailed Horizon Europe data management requirements will be announced at the end of 2020.) Data management sections of project applications are evaluated by the European Commission under the criterion 'impact.' When completing the general information section of EU funding applications, 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 cannot be made available, explain why (v) How will the data be curated and preserved?[49] The EUI Library assists project managers and principal investigators with the data management sections of funding applications and data management plans.[50]

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. The EC provides further details in the Guidelines on Data Management in Horizon 2020.[51] The DMPonline tool[52] (Section 3 above) can be used for EU data management plans, by selecting the 'Horizon 2020' template in the Funder section ('Horizon Europe' from 2021).

The European Commission's policy is for research data to be "as open as possible, as closed as necessary".[53] A list of opt-outs is provided by the EC. 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."[54]

### 6(a) European Open Science Cloud (EOSC)

The European Open Science Cloud portal was launched in 2018.[55] The EOSC portal enables the discovery of data; provides tools for data analysis; and gives information

[47] DataCite metadata schema: https://schema.datacite.org/
[48] https://ec.europa.eu/programmes/horizon2020/
[49] Guidelines on Data Management in Horizon 2020, December 2013, p.2. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[50] econlibrary@eui.eu
[51] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf p.5
[52] http://www.dcc.ac.uk/dmponline
[53] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[54] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf p.10
[55] https://eosc-portal.eu/

about storage, computation, training and security. Content and services are available via the EOSC catalogue.[56] The European Commission has published two supporting documents: Prompting EOSC in Practice[57] which explains the governance of the new service, and Turning FAIR into Reality[58] 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 initiative was launched in 2019. SSHOC constitutes the social sciences and humanities cluster of the European Open Science Cloud, and 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 Infrastructure Consortiums) and foster interdisciplinary research and collaboration."[59] There are 47 institutional participants, under the coordination of the Consortium of European Social Science Data Archives (CESSDA).

## 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[60] and Google Dataset Search.[61]

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 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(b)). Library staff can help EUI members determine whether and when a dataset can be openly shared.

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).

[56] https://eosc-portal.eu/
[57] https://publications.europa.eu/en/web/eu-law-and-publications/publication-detail/-/publication/5253a1af-ee10-11e8-b690-01aa75ed71a1
[58] https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283
[59] https://www.cessda.eu/About/Projects/Current-projects/SSHOC
[60] http://www.re3data.org/
[61] https://datasetsearch.research.google.com/

## 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 for qualitative data

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.).

Qualitative data can be generated from surveys, free-text responses to interview questions, focus group observation and/or recordings and experimental simulations. Human 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. As stated in Section 2(a): when collecting, storing, using or transferring personal data, scholars are responsible for obtaining the 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.

As with the use of micro-socioeconomic data; scholars using qualitative data should pay particular attention to ethical standards when handling data pertaining to ethnicity, health, religion, political or sexual orientation, biometrics &c. Persons, families and households cannot be identifiable in any dataset. The linking of variables to individuals, families or households is governed by data protection legislation and academic best-practice.

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.[62] 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.[63] 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

[62] http://www.eui.eu/ServicesAndAdmin/ComputingService/Software/Index.aspx
[63] http://www.eui.eu/ServicesAndAdmin/ComputingService/Software/SoftwareResearch.aspx

a 'safe-room' environment. If this is required by a data provider, contact the EUI Library.[64]

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 locked space when not in use.

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.

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 in a general repository such as Zenodo. 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.

## 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.[65] Technical support is provided at the site offices of the ICT Service.

[64] https://www.eui.eu/Research/Library/ResearchDataServices/SensitiveData
[65] http://www.eui.eu/ServicesAndAdmin/ComputingService/Software/SoftwareResearch.aspx

The EUI Library maintains the Data Portal, the Micro Data Restricted Data Server and the EUI Research Data collection in the Cadmus 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) 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.[66]

A directory of online research data software manuals, with links to full-text, is available on the Library web site.[67] 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, modelling, optimisation, regression analysis, experimental design, survey methods and time series (shelfmarks 001-006 and 500-519). Books, eBooks and manuals in any language may be suggested for acquisition by the Library.[68]

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. 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 news is also disseminated via Library Blog[69] and Twitter.[70]

## 10. INTERNATIONAL RESEARCH DATA GUIDELINES

- 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 Data Portal
- EUI Research Data outputs
- EUI Research Data submission form
- 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)

[66] https://www.eui.eu/ServicesAndAdmin/ComputingService/Software/ResearchSoftwareTutors
[67] https://www.eui.eu/Research/Library/ResearchGuides/Economics/Software
[68] econlibrary@eui.eu
[69] https://blogs.eui.eu/library/category/economics/
[70] https://twitter.com/econlibrary

- 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)
- OECD Guidelines on Research Ethics & New Forms of Data
- 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
- 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:

https://www.eui.eu/Research/Library/ResearchDataServices/Guide

econlibrary@eui.eu

Thomas Bourke

8th edition, 2020