RESEARCH DATA GUIDE

- MAY 2014
- 2.0
INTRODUCTION

Welcome to the second edition of the EUI Library's Research Data Guide, offering advice to scholars who use and generate research data. This edition contains new information about the research data pilot in the EU Horizon 2020 research and innovation funding programme (2014-2020). Feedback from the recent EUI research data survey has also been incorporated.

The Guide gives an overview of data discovery, access and terms of use, technical support, data management, describing datasets and preserving datasets.

The focus is on digital research data outputs - primarily numerical statistical data. Most of the guidelines can also be applied to other forms of research data: qualitative, textual, audio and visual - the generation of which is anticipated to increase over time.

There is a growing 'open data' trend among international organisations, researchers and government agencies to share data outputs, codebooks and software. Sections 3 to 7 cover research data management and give guidelines on open data practices.

Thomas Bourke
Economics Information Specialist

econlibrary@eui.eu

May, 2014
1. **DATA DISCOVERY, ACCESS AND TERMS OF USE**

The EUI Library maintains a Data Portal providing access to both (i) licensed data resources and (ii) open data resources of relevance to the EUI research community.

There are 43 macroeconomic and 32 micro-socioeconomic resources. The Data Portal is primarily dedicated to quantitative data resources, however the number of qualitative resources is increasing steadily.

Macroeconomic databases provide statistics on national, regional and global economic and political-economic developments from major public institutions and data publishers, including: the UN, Eurostat, OECD, ECB, Thomson Reuters, IMF, World Bank, and WIIW.

Micro-socioeconomic databases provide individual, household and firm-level data observations. Major providers include Eurostat, Bureau van Dijk, GESIS, ICPSR, UKDS and DIW. Access to micro data hosted at the EUI requires additional user registration.

Each resource in the EUI Data Portal has a dedicated resource page, providing:

- Data description
- Time period and version information
- Support links (online manuals, user communities)
- Details of access and terms of use

Open data resources are available on the internet. Major repositories are indexed in the re3data.org registry. Data accessible to EUI users under Library licence are accessible via: internet protocol (IP), networked server DVD/CD-roms, or (for micro-socioeconomic data) restricted server. The Datastream financial database is run directly from the EUI desktop ‘programmes’ menu. Data from ICPSR is downloaded by Library staff on request.

The Library also maintains a resource directory for Eurozone, EU and European Statistical Data covering both open and restricted resources.
Data users are reminded that access and use are subject to license agreements, copyright terms and data protection provisions. Open data (available on the internet) can also be subject to license conditions and restrictions on re-publication. By accessing and using data, individual users are responsible for complying with the terms of use.

Scholars who require access to data held at another facility should apply for access early in their research projects. For example, some micro-socioeconomic geo-coded data can only be accessed at the issuer's secure facility. Write to econlibrary@eui.eu for assistance.

Scholars who require access to unpublished datasets (eg. underlying data associated with a journal article) should contact the Library before writing to data owners/authors. In certain instances, it may be possible for the Library to ‘broker’ access to a dataset held at another institution.

2. SUPPORT, SOFTWARE AND INFRASTRUCTURE

Data content support and manuals are provided by the EUI Library. Software and connectivity support is provided by the EUI ICT Service.

Content support is available via econlibrary@eui.eu or by visiting the Specialist's office at the Badia Library (room 085) on weekday mornings and Tuesday and Thursday afternoons. Content support in the Economics Department (Villa San Paolo) is provided at the Departmental Library desk on Monday, Wednesday and Friday afternoons from 15:00 to 18:30.

Technical, software and connectivity support is provided by the EUI ICT Service site offices. Advice on the use of statistical software is provided by the EUI Research Software Tutors. The EUI provides access to major data software and manuals: Stata, Gauss, MATLAB, OxMetrics, R for Windows, Scientific WorkPlace, Stat/Transfer, WinEdt and WinRATS.

Every Friday during term, the Library issues an e-bulletin with statistical data news, including details of new data waves, platform upgrades, data errata and tips on how to access statistical series from Library databases and the internet. EUI members can sign up with an eui.eu account via econlibrary@eui.eu.

Data presentations are listed on the Library training page. Data news is also on the Library blog.

Research data manuals are available online. Paper copies are available at the Badia Library and the Economics Department collection (005 shelfmark range).
3. RESEARCH DATA MANAGEMENT

The management and curation of research data generated by individual scholars and by joint research projects is becoming increasingly important.

Most science funding agencies require researchers to provide ‘Data Management Plans’ (see Sections 6 and 7 below) outlining how data are described (metadata), preserved, and - where possible - made available for sharing. This is due to the growing trend among international organisations, scholars and government agencies to share data outputs.

During the life cycle of research projects, it is important to keep data safe and secure. Although laptop devices are commonly used for working with data, it is best to keep regular backups on safely-secured external memory devices and/or the network server. Note that, in accordance with contractual agreements, micro data should only be elaborated on a desktop device in a secure location.

It is also important to keep an accurate record of dataset changes, variables, characteristics, software versioning and - in the case of survey data - pre-agreed terms of disclosure for possible open data sharing via repository, portal or data catalogue. ‘Metadata’ refers to the descriptors or ‘tags’ that identify a dataset.

4. METADATA: DESCRIBING DATA OUTPUTS

The metadata fields suggested below can be used as a checklist for managing data throughout the research project cycle. This metadata scheme can be updated as the project evolves. Not all of the fields suggested below will be required by data funders.

TITLE OF DATASET
The title should succinctly convey the nature and scope of the dataset. Dossiers and files should be tagged consistently.

NAME(S) OF DATASET CREATOR(S)
The name, or names, of the researchers and technical collaborators who created/elaborated the dataset should be indicated, preferably with institutional contacts. Where researcher ID numbers are available, these should be provided (eg. ORCID).

DESCRIPTION OF DATA
Provide a meaningful ‘data abstract’, indicating the kind of data, the scope of the research context and a note on data methodology. The abstract should be updated as the research project evolves.
SOURCE OF DATA
Cite the source(s) of the data. If the dataset is a primary output of a research project, this
should be stated with details of data collection (eg. survey parameters). If the dataset is
derived from a pre-existing database; cite the publisher, author and/or institutional
source. If there are multiple streams of input - all should be cited. Data that are made
publically available for sharing should be the product of significant value-added
elaboration - not a simple platforming of pre-existing data for re-distribution.

CREATION DATE
Give the date of the iteration of the dataset. If the dataset is a new wave of a pre-existing
series, this should be clearly indicated with version designation.

SPATIAL COVERAGE OF DATASET
Indicate the geographical scope of the dataset (national, regional, global) as applicable.

TEMPORAL COVERAGE OF DATASET
Indicate the start- and end-dates of dataset coverage.

FORMAT
The software format of the data should be given (eg. Stata 13, csv, &c.), with a note on
the names of dossiers, files and (for tables) tabs.

LANGUAGE
The language of data variables, notes and codebooks should be indicated. Multi-lingual
language variables (eg. for international surveys) should be listed.

LOCATION OF DATA
Provide the name of the location, primary repository, portal or web site where the data
resides. During the research cycle, researchers should consider how their data will be
preserved on a long-term basis (eg. digital archive, subject/institutional repository or
portal). Software version changes and transfers should be noted.

DIGITAL OBJECT IDENTIFIER
Most data repositories automatically issue DOIs. If a repository does not provide a DOI,
contact the EUI Library.

ACCESS STATUS AND EMBARGO
The access status should be indicated. The status ‘Open Data’ should be assigned to
datasets that are publicly accessible on the internet. If the data is subject to embargo
(eg. two years) this should be indicated. If the data is restricted (eg. on-site consultation
only), or in a closed archive - this should also be stated.

LICENSE
The terms of access and terms of use (eg. Creative Commons license) should be
indicated.
FUNDING STATEMENT
The name of funding bodies, and research grant number, should be provided where applicable.

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

DATASET CITATION
A ‘ready-to-use’ citation reference for the dataset should be provided – incorporating the core descriptive elements.

5. DATA SHARING AND OPEN DATA
By carefully recording the metadata elements suggested above, researchers will have a ready ‘checklist’ for determining whether datasets can be made available on an open data basis for sharing with other scholars.

Not all research data outputs can be openly shared. The two most significant considerations when determining whether a research dataset can be made available on an open data basis are:

- **Data protection**: For data sets where persons, households or firms are identified - it is normally not possible to publish data openly. In such instances, it may be possible (pending terms) to generate an anonymised version of the data for sharing.
- **Database copyright**: Where a research dataset is derived/elaborated from pre-existing, commercially-licensed database(s), it is normally not possible to re-publish significant extractions due to copyright protection and contractual terms of use.

Note that, even where data are provided on an open data basis, it is not always possible to re-publish freely. Write to econlibrary@eui.eu for further information on data sharing. In addition to the sharing of raw data, it is increasingly required that researchers share codebooks and, where possible, software applications used to elaborate primary data.

A directory of research data repositories is maintained by [re3data.org](http://re3data.org). Examples of data repositories include:

- GESIS datorium
- NESSTAR
- Zenodo
- Dataverse
6. DATA MANAGEMENT IN EU HORIZON 2020

Horizon 2020 is the EU Research & Innovation funding programme, 2014-2020.

EUI members preparing applications for Horizon 2020 funding are required to submit general information on data management as part of their proposal. This will be evaluated by the EC under the criterion 'impact.' Social science and economic research projects are included in the ‘Societal Challenges’ cluster.

The EUI Library can assist project leaders with the data sections of funding applications. To make an appointment, please write to econlibrary@eui.eu.

Project managers should address the following questions when completing the general information section of the application:

- What types of data will the project generate/collect?
- What standards will be used?
- How will this data be exploited and/or shared/made accessible for verification and re-use? If data cannot be made available, explain why.
- How will this data be curated and preserved?1

In the context of Horizon 2020, the EC has launched a research data pilot. For those projects participating in the research data pilot, a more detailed data management plan (DMP) is mandatory within six months of grant confirmation. The EC anticipates that project DMPs will evolve throughout the research cycle.

The EC's DMP template is on p.5 of Guidelines on Data Management in Horizon 2020 (version 1.0).

Grant beneficiaries must "deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate - free of charge for any user - the following: (i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; (ii) other data, including associated metadata, as specified and within the deadlines laid down in the data management plan.” (EC Guidelines, p.3.)

7. INTERNATIONAL RESOURCES AND GUIDELINES

Further information is available from these sources:

- Open Research Data Pilot in EU Horizon 2020
- EC Report on Open Research Data
- GESIS Datorium
- Göttingen/OpenAIRE Study on the Protection of Research Data
- Digital Curation Centre guidelines on preserving and sharing research data
- Open Economics Principles from the Open Knowledge Foundation
- MANTRA data advice for scholars
- EU Open Data Portal
- UK Data Service guidelines on managing data
- National Science Foundation guidelines for social science and economic data plans
- American Economic Association guidelines on data availability
- re3data.org registry of international data repositories
- Zenodo data repository (CERN & OpenAire)
- RECODE project on open data strategy for Europe
- DLA Piper rights in data handbook
- U.S. Government open data policy
- A directory of research data repositories is on this Library page

8. EUI LIBRARY DATA SERVICES

The EUI Library provides advice on the following aspects of data curation:

- Metadata for research datasets
- Ontology (eg. descriptive ‘tree’ for multi-level datasets)
- Data preservation
- Data sharing, guidelines for open data, and limitations
- Data documentation

Please consult the Library Data Services page or write to econlibrary@eui.eu for assistance.