Data publication:
Principles and practices of open data sharing

Michael Murigi
Monitoring and Evaluation Coordinator, PEP

www.pep-net.org
Why share data?

- **Research data is a public good**
  - Should be accessible for re-use by other researchers

- **Other researchers may use your data** to:
  - Contextualize your published article
  - Replicate your study
  - Develop their own research based on your data
  - Incorporate your data into a larger dataset or use multiple datasets together

- **Increases visibility** of you as a researcher and of your research work
Data sharing options

- Email, on request
- Supplementary files
- Personal/institutional website
- Trustworthy data repository
- Data paper
• Access does not mean permission to use

• Provide users with rules/guidelines on how your data can or can’t be used
  o Institutions/countries/regions where data can be re-used
  o Persons that can re-use the data (bona fide researchers versus private citizens)
  o Permit (or limit) commercial re-use

• Users may implicitly agree to terms of use when accessing downloading the data from repositories or explicitly e.g., to a pop-up.
A license agreement is applied to data deposited in a repository.

Creative common (CC) licenses – a standard format for licensing content.

Types of CC licenses:
- CC Attribution (BY) – mandates attribution (e.g., citation)
- CC Non-commercial (NC) – for any non-commercial use
- CC No derivatives (ND) – prohibits production/sharing of derivatives
- CC Share-alike (SA) – any purpose if derivatives licensed under same terms
Metadata

- “Data about data”

- Contextual and descriptive information

- Examples of metadata:
  - describing the dataset (title)
  - about the dataset’s creators
  - describing the contents of the dataset
  - about the dataset’s related publications
<table>
<thead>
<tr>
<th>Typical research journal</th>
<th>Typical data repository</th>
<th>Typical data journal</th>
</tr>
</thead>
</table>
| (a) Dataset checking may be available as additional paid service | Data files reviewed by repository staff for  
• Technical integrity of data  
• Violation of repository conditions of use | Data curation and checking offered as part of the normal data journal workflow |
| (b) Reviewers may be allowed to request to see data | May facilitate peer review of data prior to it being publicly available | Data peer reviewed |
| (c) Data referencing may/not be allowed | May provide guidance to data users on how data should be cited | Journal staff typically check if data have been correctly cited |
| (d) Final output is a citable scholarly article | Final output is a citable dataset | Final output is a citable data article |
Deposit data in an appropriate repository

Draft a data descriptor/manuscript based on a template

Submit your manuscript

Peer review of the manuscript and the data (by data curators)

Revise the manuscript and data, as required

The data descriptor/paper is published

Process: Publishing your data paper with a data journal
Data descriptor/manuscript template: Contents

- **Title** of the dataset
- **Authors** and their affiliations
- Abstract
- Background
- **Methods** of data collection
- Data record
- Competing interests
- References
<table>
<thead>
<tr>
<th>Repository name</th>
<th>Information on fees/costs</th>
<th>Size limits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dryad Digital Repository</strong></td>
<td>$120 USD for first 20 GB, and $50 USD for each additional 10 GB</td>
<td>None stated</td>
</tr>
<tr>
<td><strong>figshare</strong></td>
<td>100 GB free per <em>Scientific Data</em> manuscript. <a href="https://www.figshare.com/about/faq#pricing">Additional fees apply for larger datasets</a></td>
<td>1 TB per dataset</td>
</tr>
<tr>
<td><strong>Harvard Dataverse</strong></td>
<td><a href="https://dataverse.harvard.edu/about">Contact repository</a> for datasets over 1 TB</td>
<td>2.5 GB per file, 10 GB per dataset</td>
</tr>
<tr>
<td><strong>Open Science Framework</strong></td>
<td>Free of charge</td>
<td>5 GB per file, multiple files can be uploaded</td>
</tr>
<tr>
<td><strong>Zenodo</strong></td>
<td><a href="https://zenodo.org/about">Donations towards sustainability encouraged</a></td>
<td>50 GB per dataset</td>
</tr>
<tr>
<td><strong>Mendeley Data</strong></td>
<td><a href="https://mendeley.com/about">Contact repository for datasets over 10 GB</a></td>
<td>10 GB per dataset</td>
</tr>
</tbody>
</table>
## Relevant data journals for social science data

<table>
<thead>
<tr>
<th>Data journal</th>
<th>Article processing charges</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Journal of Open Humanities Data</strong></td>
<td>£ 100 per data paper</td>
</tr>
<tr>
<td><strong>Scientific Data</strong></td>
<td>100 GB free per Scientific Data manuscript. Additional fees apply for larger datasets</td>
</tr>
<tr>
<td><strong>Elsevier Data in Brief</strong></td>
<td>Not stated</td>
</tr>
<tr>
<td><strong>Journal of Big Data</strong></td>
<td>£1060.00/$1570.00/€1290.00 per article</td>
</tr>
</tbody>
</table>
Data citation and referencing

• Why are data citation and referencing important?
  o Provide credit to those who created or generated the dataset.
  o Support reproducibility
  o Increase the citation rate of the paper.

• Data referencing style:
  Authors/creators > Title > Publisher (Repository) > Persistent identifier > Publication year

Impact metrics

Useful to track re-use of data
Including, number of:

- Times data have been viewed
- Times data have been downloaded
- Times data have been cited in a scholarly publication
- Newspaper articles written about a dataset
- Policies referencing/making use of a dataset