Recommendations for Research Data Management (RDM)

1. Considerations for the Preparation of a Data Management Plan

Qualitative and quantitative scientific research generates a significant amount of research data[1] which researchers want to keep safe and for which they are responsible in terms of security, storage and preservation. Therefore, it is necessary to set up a data management plan which is a 1–2-page document, tailored to the different stages of the research and describing in detail how the data generated will be managed during the whole research process.

When preparing research plans, it is becoming increasingly common for research funders or funding organisations (e.g., OTKA, Élvonal proposals, Horizon 2020) to request that a data management plan be included within the research concept, thus promoting thoughtful data management, accessible, sustainable, and reusable data, knowledge sharing and more active participation in open science. The principle of *"as open as possible*, *as closed as necessary"* should be considered and taken into account when providing access to data and datasets.

In fact, a data management plan should be outlined at the very beginning of the research, as early as the planning stage. This plan should include the most important information on data management:

- how and for what purpose data will be collected,
- who will be responsible for the data (i.e., who will be the owner of the data),
- what types of data will be generated,
- size of the dataset, what and how much storage space will be needed,
- how it will be managed, where it will be stored,
- how to ensure information security and data security,
- what happens to data after the research is completed,
- to whom your data may be relevant for re-use,
- standards, methods and metadata: reasoning you applied specific methods, standards to collect and manage your data,
- describing the rights to access your database (using a licence: when, how and who can access them),
- where necessary, ensure the protection of personal data, including anonymity.

[1] Research data are data collected and recorded in a non-digital (paper-based) or digital format during the research.

Although there are some important common points in the data management plan (some proposals use separate forms, such as OTKA proposals mentioned above) which the researcher must answer, these RDMPs may differ significantly from each other, reflecting the specific characteristics of the particular research project.
Managing your research data should follow the **FAIR Principles**[2]. Therefore, your archived data should be:

F: findable by others and therefore properly supplied by detailed metadata (globally unique persistent identifier),

A: accessible through an easy access service (standardised communication protocol),

I: Interoperable (data should be readable by machines and humans without specialised algorithms, Metadata formats should use shared vocabularies)

R: reusable by others, with appropriate data use permissions (clear licenses and other conditions).

2. Data Security, Data Repositories

Open data sharing (in particular through the use of *data repositories*) allows research results to be exploited much more quickly, efficiently and widely, as they can be accessed by other members of the research community (in a pre-defined way, controlled by the researcher or funder), which facilitates knowledge sharing.

The safest way to store research data is to upload it to a *data repository*, which helps to ensure that your data is secure in the long term by providing a number of data security requirements.

This requires you to think about your data management strategy, how to manage your data throughout the research, how to store it physically and virtually, how to protect it and ensure that only authorised people have access to it, and how to share it with the wider scientific community after the research is completed.

The three most important data security requirements to keep in mind when choosing where to store data are:

- security: data and information content should only be accessed by appropriately authorised individuals, when they are permitted to do so (e.g., after a period of time, after an embargo has expired), and only at the level to which they are authorised by us. The same applies to the use or possible re-use of data,

[2] The FAIR guiding principles proposed by the EU and scientific communities promote Open Science and support open access and reusability of research data.

- integrity: data must not be modified by unauthorised persons, and you must ensure that it remains in its original state,

- availability: data, information and related infrastructure should be available where and when needed.

2.1. Data safety storage options

It is important to store your data under secure conditions on at least 2-3 physically separated locations to avoid data loss. The safest option may be uploading your data to an on-line data repository.
Data and information generated during the research are usually managed and stored on one of the following devices or channels, therefore data and information security should also apply to the following areas:

- the paper on which it is recorded,
- the hardware on which it is stored and processed:
    - own PC/notebook drive (HDD, eMMC),
    - flash drive, DVD,
    - institutional drive, server,
- the cloud service where it is stored (e.g., Google cloud),
- on-line data repository storage, cloud (e.g., Zenodo),
- the software used to process it,
- the operating system that runs the software,
- the communication channel over which the data flows,
- the users who manage data in general.

2.2. Cloud-based data repositories

Hungarian data repositories *(still in test function)***:**

- MTA SZTAKI data repository (CONCORDA): https://science-data.hu/ (formerly: https://concorda.sztaki.hu) originally set up on an experimental basis to collect Covid-19 research, but now any database can be uploaded.
- University of Debrecen data repository: https://adattar.unideb.hu/.

International interdisciplinary repositories:

- Zenodo (OpenAIRE): https://zenodo.org/
- Dataverse: https://dataverse.org/
- Dryad: https://datadryad.org/stash
- FigShare: https://figshare.com
- Mendeley Data (Elsevier): https://data.mendeley.com ' research-data
- Open Science Framework (OSF): https://osf.io
- Globus: https://www.globus.org
- ELN'S (Labguru, Labarchives): https://www.labarchives.com/eln-for-research/

Of the above, the most popular free repository with appropriate data security requirements is *Zenodo*[3], a general-purpose, open-access data repository developed in the framework of the European OpenAIRE programme, where you can upload up to 50GB/database. Researchers and institutions can also upload their databases related to their publications. There are also other data repository search engines which are specifically designed to help you choose the most appropriate repository for your purposes and data:

- https://www.re3data.org/search
- http://oad.simmons.edu/oadwiki/Data_repositories
- https://www.nature.com/sdata/policies/repositories
- https://fairsharing.org/databases
- https://repositoryfinder.datacite.org
- https://explore.openaire.eu/search/find.

3. Data protection, ethical aspects

The protection of personal data[4] is of paramount importance. Any data or information that can be used to identify a person, family or household member is considered personal data, and you have to protect it as it is subject to the *Hungarian Data Protection Regulation*[5] and, since 2018, you have to comply with the GDPR[6] *(General Data Protection Regulation)* as well.

In the research data management plan (RDMP):

- you must include how you will ensure the protection of your data at each stage of the entire research, e.g., how you will protect documents; how you will store data physically and virtually, how you will ensure that they are not accessible to others (e.g., wether you will store your questionnaires and personal data physically in a different room, or in a safe, or whether you will provide your digital data with a virtual crypt

[3] The Zenodo data repository is operated by the CERN Data Centre.
[4] Examples of personal data: surname and first name, address, ID card number, location data (including virtual data such as IP address), mobile phone number etc.
[5] Act LXIII of 1992 On the Protection of Personal Data and the Publicity of Data of Public Interest.
[6] The GDPR (*General Data Protection Regulation*) is regulation No. 2016/679 passed by the European Parliament and the Council, which entered into force in Hungary in May 2018 and protects the data of persons and provides regulation for the free flow of information between member states.

code, etc.)

- you must also *anonymise* (in a reversible way) your data (e.g., by using some kind of cryptographic system). A simple and compliant anonymisation process is facilitated by an anonymisation algorithm called Amnesia, available online at https://amnesia.openaire.eu/amnesia;
- take into account the relevant rules of the GDPR, e.g., if you fill in a questionnaire, you must state that the data will be treated confidentially and used only for the original research purpose;
- state how you will ensure that the respondent cannot be identified at individual, family or household level when processing the data (e.g., using unique identifiers, a random number generator, aggregated data, etc).

4. Data management plan (RDMP) in practice

There are open source tools available on the Internet to help you prepare a data management plan online, which will guide you through all the main points and provide you with a personalised template to help you prepare your data management plan – unless your research funder requires the use of a specific form. Examples of such effective tools include:

- DMPTool[7]: https://dmptool.org/

- DMPOnline[8]: http://dcc.ac.uk/dmponline/

The data management plan should include:

- information about the data: what type of data you collect, for what purpose, how this relates to the original objectives of your project,

- what the format of the data is (e.g., tables, records, images),

- what the size of data will be (e.g., MB, TB, ZB size), how much and what storage space you will need (e.g., consider whether you will store your data as a hard copy or only electronically),

- how the data will be handled, stored and used, from data collection through the different phases of the research and after completing the research,

- whether there will be links to other public databases/data repositories,

- secondary re-use information: further intended use of a specific dataset,

- clarify data accessibility, reuse, sharing, licensing: plan for data sharing, how the data will be made available, how it can be used by others (e.g., public data repositories (see 2.2) or sent on request, recording access mechanisms where appropriate). This may be the researcher's or research team's own decision, but often it may also be influenced by the client, funder,

[7] DMPTool is produced by the University of California Curation Centre of the California Digital Library.
[8] DMPOnline is a product of the UK Digital Curation Centre.

contracting authority and other aspects(e.g., the position of consortium members). If you do not wish to share your data at all, specify the reasons for not doing so.

- similarly, it should be clear how and when metadata will be accessed, by whom, and where and how it will be stored;

- what metadata[9] standards will be used (DublinCore[10], Datacite[11], etc.);

- data ownership rights: any restrictions on data sharing due to the need to protect data, metadata, who will supervise the repository if required;

- a timeframe for making the data public, or the setting of an embargo period for making the data available to the public, or a timetable for making the data public;

- the format of the final database (e.g., lab notebook, jpg, or in the case of quantitative social science research, most often a file with the extension cvs, sav, sys or dta, do);

- consider what the costs of long-term storage of your data might be (e.g., the cost of anonymisation procedures, DOIs, etc.), and include these costs in your data management and budget plan;

- what permanent or unique identifiers (e.g., DOI[12] or ORCID[13]) will be used for the database;

- it is worth revising[14] the data management plan from time to time during the research and updating it if necessary, as aspects may change, new innovation policies, ICT technology requirements, infrastructures and so on, may emerge.

[9] **Metadata** is *"data about the data"*, for a document or database it contains the most important descriptive properties, author, year of publication, etc. This data can greatly help in finding documents and data.
[10] **The Dublin Core Metadata Initiative** is an internationally accepted method for standardising the metadata of documents available online, and which makes it easier for search engines to access documents by creating a digital "*library card catalogue"*.
[11] **DataCite** is a non-profit organisation that aims to facilitate the on-line availability and discoverability of research data and results for its members through the use of DOIs or other identifiers.
[12] **DOI *(Digital Object Identifier)*,** registered with the **CrossRef** agency, is a unique identifier that helps to make scientific publications available in on-line format.
[13] **The ORCID *(Open Research and Contributor Identifier)*** is an international author identification code that collects researchers' publications based on DOI identifiers and helps to identify the researcher.
[14] Including a time-table in your RDMP may help us revise your document from time to time.