Research Data Management

A Guide for Researchers

Research Data Working Group
in the Priority Initiative “Digital Information” of
the Alliance of German Science
Organisations

This document is intended to provide basic guidance to researchers who work with digital data as well as all stakeholders with an interest in this issue and also provides advice on sources of further information. It was prepared by the Research Data Working Group in the Priority Initiative “Digital Information” of the Alliance of German Science Organisations[1].

[1] http://www.allianzinitiative.de/en/

Definition and Lifecycle of Research Data

The Priority Initiative "Digital Information" of the Alliance of German Science Organisations defines research data as "…data produced by scientific projects, for example by means of digitization, the study of sources, experiments, measurements, surveys or questionnaires".[2]
Research data management refers to the process of administering this data throughout its lifecycle, from planning, production, selection and evaluation to storage and processing for the purposes of reuse. Due to the heterogeneity of data, it is necessary to develop solutions tailored to specific disciplines or in some cases specific types of data.

Not all activities that form part of this process necessarily have to be performed by researchers themselves. The standardisation of individual processes in the data cycle reduces the workload on researchers in the course of their project work. It also allows staff and scientific resources to be made available for certain tasks, such as the documentation, storage and evaluation of data, and to provide central services of this type for all researchers.

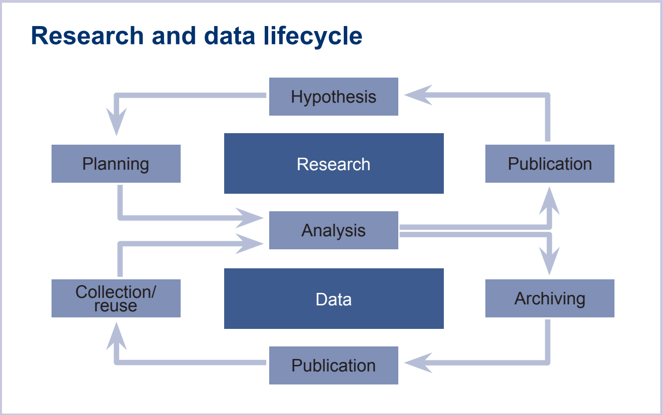

Fig. 1: *Research and data cycle in science and the humanities*

[2] http://www.allianzinitiative.de/en/

Why Implement Research Data Management?

In many scientific disciplines, the management of research data is not (yet) mandatory. As shown in Fig. 1, this process comprises a range of measures from the planning of data gathering to data management, data publication and concepts for deletion. This requires well-trained personnel and demands both time and money. In many cases, researchers are also reluctant to make their data available for reuse. As well as anxiety about transparency creating vulnerability, researchers cite reasons including the costs involved (while others can 'help themselves' for free), a lack of control over what happens to the data when it is reused, and a lack of awareness as to where the data can be provided. In addition, there is often uncertainty regarding copyright issues.

These concerns are countered by an array of benefits:
To comply with the rules of good scientific practice[3] it is already the case that the data generated during the course of a project must be archived after the project is completed in such a way that it retains its meaningfulness. This data is an important component of the scientific value creation chain. Quality-assured data stored in a long-term archive that is described with metadata and to which there is unrestricted access has considerable medium- and long-term benefits for research. It eliminates the need to repeat investigations and experiments, with the associated time and costs. Making data generally available increases the transparency of research, allows reliable quality control and, last but not least, allows researchers to validate their own results. From a long-term perspective, research data management allows future methods to be applied to existing data and thus provides greater confidence as to the value of investment in current projects.

Data is considered a 'research product', which is increasingly exploited for evaluation purposes. Appropriately stored and presented, it can be cited just like research papers, thus enhancing the academic visibility and reputation of both the individual researcher and his or her institution. It has been shown that publications that appear along with the relevant research data are cited more frequently than papers without this ac-

[3] http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

companying evidence.[4] Finally, effective research data management can allow research results to be made available promptly on recognised platforms such as repositories, and thus put to use in practice more rapidly (the process of translation). By providing quality-assured research data as part of an efficient, structured process, research institutions can position themselves as active players in the national and international scientific environment and thus make an important contribution to promoting networking between institutions and researchers.

Many stakeholders, such as research institutions and research funding agencies, have established mandatory requirements relating to research data. Information about this can usually be found in an institution's or funding organisation's data policies. Many scientific publishers now also require authors to provide the research data associated with a publication. This is governed in contracts between author(s) and publisher. Third-party funding bodies such as the German Science Foundation (DFG), the German Federal Ministry of Education and Research (BMBF) and the European Union, through the Framework Programme for Research and Innovation Horizon 2020, offer the option of applying for funding specifically for the purposes of research data management within a funded project. Details are provided in the relevant calls and guidelines, which usually also specify a requirement for data management plans.

[4] Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Data Management Plans

A data management plan (DMP) is an instrument for the structured documentation and processing of all relevant information relating to the data used in a research project. It allows any problems to be identified at a very early stage, for example, the need for declarations of consent in the case of questionnaires or more advanced hardware requirements in the case of computer simulations. It is also a useful way of giving the various project participants a clear overview of how data will be handled within the project. It makes sense to prepare the DMP in parallel with the planning of a project and it should be updated during the project in response to changing circumstances. Increasingly, funding organisations also require grant recipients to draw up DMPs in order to document the management of project data. In this context, a valid DMP can help to explain the reasons for additional data management costs and enable applicants to secure the necessary resources.

As a general rule, a DMP should answer the following questions:

- What kind of data will be collected, how, in what quantity and in what format?
- How and where will the data be stored and what additional information (metadata) will be required?
- What legal, ethical or technical requirements must be satisfied in the gathering, storage or (if relevant) dissemination of the data?
- Is the data to be published after completion of the project and if so, where and with what conditions?

A DMP can be drawn up with the aid of specialist software or online tools. For example, the European Commission recommends the tool offered by the UK's Digital Curation Centre.5 Tools specifically adapted to German research and funding structures are currently at the development stage (e.g. RDMO6). Many research institutions have also already set up central departments to provide researchers with advice and support on data management and preparing a DMP.

[5] https://dmponline.dcc.ac.uk/
[6] http://rdmorganiser.github.io/

Storing Research Data

There are various options available for the storage of research data. It can be stored in data archives or data repositories. The first step is to find out whether your institution operates such a platform (an institutional data archive or repository) or collaborates with a data centre where it may be compulsory to store your data. If this is not the case, it is advisable, particularly with a view to allowing other researchers to reuse your data, to find out whether any subject-specific repositories exist for your particular discipline where your data can be stored. A useful tool for finding subject-specific repositories is Re3data (Registry of Research Data Repositories[7]), which currently lists over 1,500 repositories in all research areas. Some repositories offer free access to data while others offer restricted access or no access rights at all. Zenodo[8] is a repository for research data from all fields of research. Zenodo was developed by CERN as part of the European Commission's OpenAIRE project[9] and gives all researchers the option of storing their data and making it available on a long-term basis. Dryad[10] is a repository specifically for data on which a publication is based, which is now required by nearly all scientific publishers in the form of 'supplementary material' when a paper is submitted.

Because research data is generally of high value, the choice of data repository should be made very carefully and with due consideration of the specific interests involved. The suitability of commercial cloud services for data storage will require careful consideration on a case-by-case basis. Depending on individual requirements, the criteria to consider may include:

[7] http://www.re3data.org/
[8] https://zenodo.org
[9] https://www.openaire.eu/
[10] https://datadryad.org/

- Who operates the repository? A major publicly funded organisation may be a safer choice than a commercial provider, particularly where long-term availability is concerned.
- Is there a suitable choice of different licensing models defining how data can be reused?
- Does the repository enable collaborative work with the data?
- Do charges apply for data storage?
- Is the data recorded and processed in a professionally appropriate way? For example, is there a good search facility and is it easy to cite the data, for instance through the assignment of a DOI[11]?

[11] See “Data Citation”, p. 11

Open Data

The Berlin Declaration on Open Access[12], which was signed by all the leading research organisations in 2003, not only covers open access to published research but also addresses the question of research data. The overall value of research data to the scientific research process means that a growing number of funding organisations are introducing a requirement for open access to the data generated in projects that they fund. For example, in Horizon 2020 – the Framework Programme for Research and Innovation – the European Commission integrated an Open Research Data Pilot“[13]. Here, funding recipients are required to make the data generated during funded projects freely accessible and reusable.

For researchers, open access to their data offers the benefit that it can be reused and therefore cited by others. But there are also good reasons for not publishing research data. Obviously, not all data can legitimately be made freely available. In some circumstances, it may be necessary to comply with the requirements of project partners or legal requirements, for instance data protection regulations concerning personal data. So before storing data in a repository, it is important to consider carefully whether access should be open, restricted or not permitted. For cases such as these, Open Research Data Pilot offers an opt-out. This means that in justified exceptional cases, contrary to official requirements, research data does not have to be made freely available.

A good way to define how others can reuse your research data is a Creative Commons Licence[14]. Data producers can use these licences to define under what circumstances their data can be reused.

[12] https://openaccess.mpg.de/Berlin-Declaration

[13] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[14] https://creativecommons.org/

Data Citation

In keeping with the recommendations for upholding good scientific practice, work from other sources, that is not self generated, must be cited.. As well as information from scientific publications, it is now also considered important to cite research data. Citation ensures that the work done by the data producer is acknowledged. Citations of research data are already being counted and made visible in dedicated databases and bibliometric analysis tools (e.g. the Data Citation Index and tools based on alternative metrics such as Altmetric or Impact Story). To ensure reliable and long-term citation, it is recommended that data should be assigned a persistent identifier. For research data, a DOI (Digital Object Identifier) is often used. In some repositories this is generated automatically when a data set is uploaded. Alternatively it can be assigned by the institution that publishes the data set. If in doubt, you should enquire at your own institution’s library as to whether the library or another internal department can assign DOIs.

Data Journals

If you take the time to add explanatory metadata to your high-quality research data and then publish your data, you will want to receive the appropriate recognition from the community. One way of achieving this, in addition to the aforementioned data repository option, is the increasing popular method of publishing data papers in mixed or data-only journals.

A data paper does not contain scientific findings but is a simple description of the data, the methods with which it was generated and, if relevant, how the data was processed. The paper will, of course, reference the data itself, which is normally stored in an external data repository (see Storing Research Data). The data paper can be referenced from research publications and serves as a semantic link between publication and data set. It can be cited by researchers who wish to reuse the data and thus contributes to the established reputation mechanism. Data papers, like conventional research papers, normally also go through a peer review process. The vast majority of data journals are published on an open access basis, i.e. there are no subscription costs, although there may be Article Processing Charges (APC) for publication.

Finding Research Data

The more extensive and varied the amount of high-quality, available research data, the more attractive it becomes to reuse this data for your own research. While in some disciplines, such as genetics, data is stored in a small number of central archives[15], in other subject areas data is distributed widely among different repositories worldwide. As yet, platforms that bring together information about this data are rare. However, there is the commercial Data Citation Index as a module of the Web of Science as well as the metadata search offered by DataCite[16], the umbrella association of organisations that issue Digital Object Identifiers (DOIs) for research data. National initiatives to make research data accessible through standardised services already exist in Australia[17] and th UK[18]. At European level, CESSDA[19] brings together research data in the social sciences and Europeana[20] in the humanities. Some research institutions also offer central services of this nature which help users to locate particular data.

[15] http://www.ebi.ac.uk/

[16] http://search.datacite.org/

[17] https://researchdata.ands.org.au/

[18] http://ckan.data.alpha.jisc.ac.uk/

[19] http://cessda.net/

[20] http://www.europeana.eu/

About this Publication

The online version of this publication is available at:

http://doi.org/10.2312/allianzoa.030

Published by
Research Data Working Group in the Priority Initiative “Digital Information” of the Alliance of German Science Organisations

Edited by
Michael Franke (Max Planck Digital Library)
Henriette Senst (Robert Koch Institute)
Dagmar Sitek (German Cancer Research Centre)

Contact
Dr. Stefan Winkler-Nees (DFG)
Stefan.Winkler-Nees@dfg.de

Version dated
1st edition, April 2018

DFG Deutsche Forschungsgemeinschaft

Fraunhofer

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HRK Hochschulrektorenkonferenz
Die Stimme der Hochschulen

Leibniz
Leibniz-Gemeinschaft

Leopoldina
Nationale Akademie der Wissenschaften

DAAD Deutscher Akademischer Austauschdienst
German Academic Exchange Service

MAX-PLANCK-GESELLSCHAFT

WR WISSENSCHAFTSRAT

© Vege / Fotolia