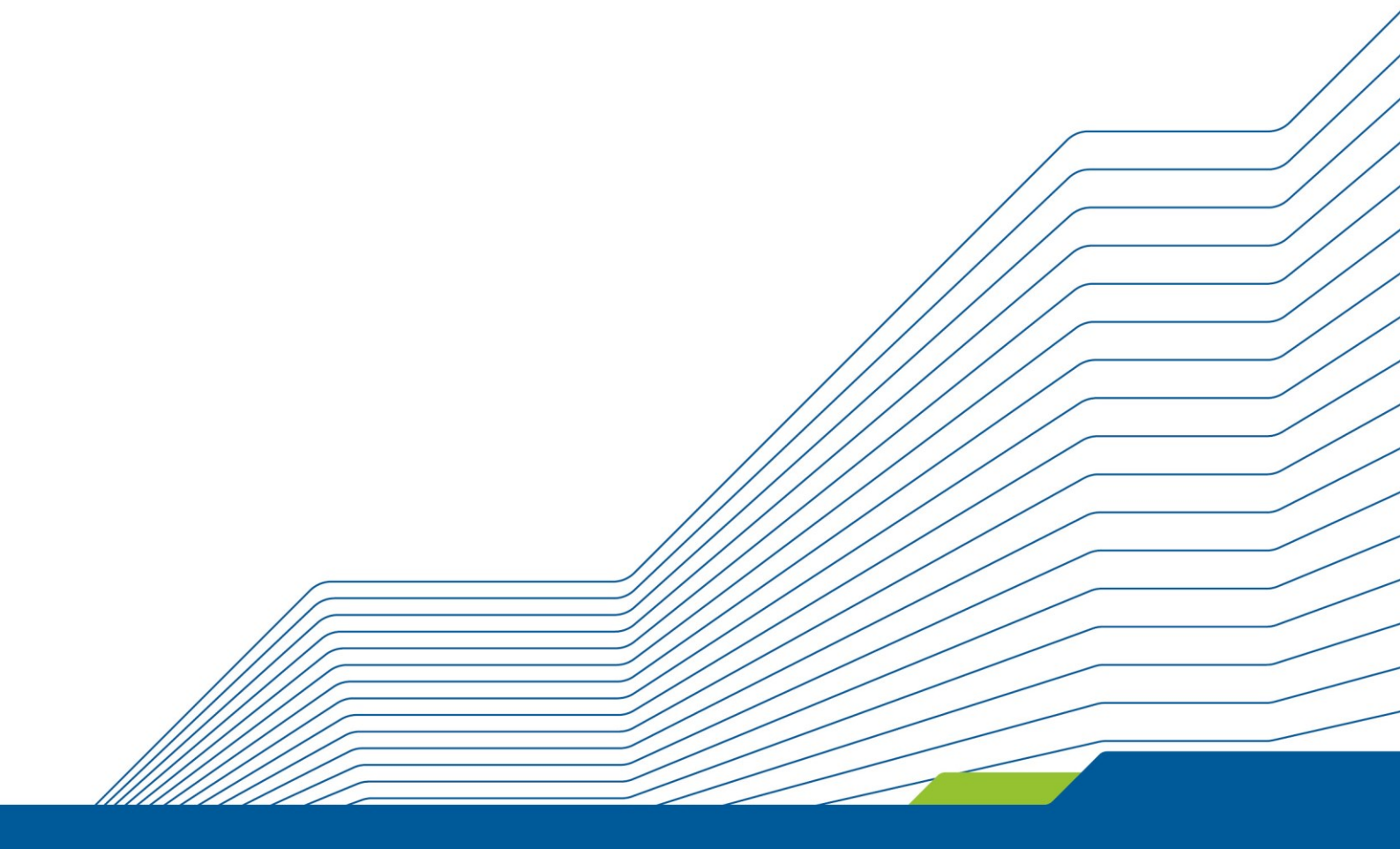
HELMHOLTZ
Open Science

Recommendations for guidelines at the Helmholtz centres on handling research data

Approved at the 109th General Assembly of the Helmholtz Association on 13-14 September 2017.



Impressum (German legal notice)

You will find the online version of this publication at:
DOI: https://doi.org/10.2312/os.helmholtz.004

Publisher

Open Science working group of the Helmholtz Association

Editors

Helmholtz Open Science Coordination Office

Contact

Helmholtz Open Science Koordinationsbüro
c/o Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum GFZ
Telegrafenberg, 14473 Potsdam
Email: open-science@helmholtz.de

Valid from

February 2019

HELMHOLTZ
Open Science

Table of contents

Preamble

These recommendations for guidelines at the Helmholtz centres on handling research data were approved by the General Assembly of the Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V., or "Helmholtz Association", on 13/14 September 2017. It was emphasised that Open Access and technology transfer were goals of equal importance for the Helmholtz Association, the potentially divergent requirements of which should be balanced against each other in individual cases.

Objectives

Due to rapid digital developments and increasing volume of data, the way research data and its use are handled is changing fundamentally. For this reason, the Helmholtz Association, at its General Assembly in September 2016, approved the document: "Make information as a resource more easy to use! A position paper on handling research data in the Helmholtz Association"[1] (below: "position paper"). In this paper, the General Assembly requests that the centres set up corresponding guidelines for handling research data and commissioned Mr. Marquardt in his function as a supporting member of the Open Science working group, in coordination with the working group, to introduce the operationalisation of the paper.

This document is directed to the relevant individuals in the centres who are responsible for upholding the basic principles approved in the position paper.[2]

The aim of these recommendations is therefore to support the centres in formulating their respective guidelines. These guidelines should be made publicly accessible to all interested parties. Here, it is taken for granted that the implementation of the recommendations presented in this paper lies within the area of responsibility of the centres and that it must be adapted to the respective specific research and disciplinary conditions. These recommendations contain suggestions as to what should be regulated in the centre-specific implementation, without specifying what form the individual aspects should take.

A key goal of these recommendations is also to achieve a certain degree of harmonisation of the guidelines on handling research data with the Helmholtz Association centres, and to help create a "Helmholtz standard". It is also necessary to develop a quality standard of this nature in light of the specific procedures for the different disciplines, standards in the different fields of research and publication bodies and the requirements of large research funding organisations. In this context, the *Guidelines on FAIR Data Management in Horizon 2020*[3] issued by the EU Commission and the *Leitlinien zum Umgang mit Forschungsdaten* by the DFG[4] should be mentioned which already require information

[1] Original title "Die Ressource Information besser nutzbar machen! Positionspapier zum Umgang mit Forschungsdaten in der Helmholtz-Gemeinschaft". Created by the Open Science working group.

[2] See also chapter "Responsibility"

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

[4] DFG Leitlinien zum Umgang mit Forschungsdaten (DFG guidelines on handling research data (30 September 2015) https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

or clear plans when an application is submitted as to how data related to these publicly funded research projects is to be handled, and how it can be processed for future use.

The implementation of these recommendations is intended to make scientists' work easier by not only guaranteeing the fulfilment of the respective requirements of the publication bodies and funding organisations by means of a common "Helmholtz standard", but also ensuring that the necessary framework conditions at the centre are created and upheld.

The Helmholtz Association sees its mandate as being to attain the social goal, among others, of obtaining quality-assured research data and rendering it usable for society in the long term. The centres should recognise the necessity of the additional effort involved in relation to research data management with their guidelines, and to support its implementation both in structural and financial terms, as well as through scientific recognition.

Research data and research data management

Here, the term **research data** is used to describe all data that is created during a research process, which is used for this process or which is a result of it. It is generated using different methods, or procured, or gathered, observed, simulated, derived, validated, processed, further developed, analysed and finally archived in a manner depending on the specific **research topic**. Accordingly, research data occurs in every scientific discipline and in different media types and formats, aggregation and quality levels depending on the stage of the data's life cycle.

For the **subsequent use of research data**, it is necessary to record metadata and document the context in which it was created and the tools or software used. In this context, long-term access to and the subsequent use of scientific software is essential.[5]

[5] The challenges, opportunities and framework conditions that arise from this are being discussed in the "Access to and subsequent use of scientific software" task group, which is part of the Open Science working group.

The term **research data management** describes the handling of research data, from the planning, its generation and processing through to long-term archiving or deletion, taking into account the regulations relating to the securing of good scientific practice. Accordingly, research data management relates to the entire life cycle of research data. Furthermore, the term also incorporates the documentation of the processes in the context in which they are recorded, specific to the discipline. **Data management plans** make it easier to document these processes and to describe the data.

In the context of the specific discipline and in light of the special features of the respective research institution, suitable regulations must be defined for the centres that determine the scope and degree of detail of the data to be gathered, as well as the specific area of application of the guidelines. This is necessary solely for the reason that effective research data management cannot function without resources that are provided specially for the purpose.

On the basis of the process described in the Horizon2020 programme, we recommend that a compact data management strategy should already be submitted in applications or plans for research activities, research projects or research infrastructures (i.e. external funding applications, as well as all Helmholtz processes and internal processes within the centres, through to programme-oriented funding and strategic investment), and during the further course of action, to make the maintenance of a detailed data management plan, which describes the handling of the research data arising within the framework of the project procedure or project, a mandatory task. Already when applying for a research proposal[6], the resources and copyright, usage and access rights, as well as storage during and after the end of the research activity, should be taken into consideration. Here, the projects should be given the opportunity to use a standard plan by the centre or an organisation unit.

Responsibility

The responsibility for managing the data of all research must be clearly regulated in the centres – up to the level of the research activity, research projects and research infrastructures. In coordination with the scientists involved, and taking into account the relevant standards issued by the research funders and partners, all stages of the research data management procedure are determined, documented and implemented accordingly. Here, ethical, legal and usage aspects are taken into account, as are adherence to the rules of good scientific practice.

[6] Research activities are here also understood as meaning projects and infrastructures.

Open Access

In agreement with international organisations in the field of funding and implementation of research tasks, the Helmholtz Association supports the long-term securing and fundamental Open Access to research data from publicly funded research in accordance with the basic principles of Open Access as set down in the "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" of 2003. Here, statutory regulations, in particular the protection of personal data, as well as the scientific interests and contractual agreements with cooperation partners and, if relevant, usage interests, should be taken into account. It should also be taken into account that under certain circumstances, the at least temporary confidentiality of research data is a mandatory requirement for later commercialisation. Reasonable embargo periods for exclusive initial use can be set according to the centre's standards and should then be made public as a part of the guidelines on handling research data at a centre.

When determining the duration of this period, the centres should orient themselves to corresponding specifications made by specialist associations, major research groups and research funders. Currently, there are no such specifications for many fields of research. The same applies to the determination of the point in time at which the period of validity should begin. The ongoing uncertainty should not lead to a failure to set a defined period. To a far greater extent, the centres should be actively involved in the creation process and check the appropriateness of the decisions made at regular intervals.

The decision in favour of a generally open way of handling research results does not mean that they cannot be commercialised. In agreement with the mission of the Helmholtz Association in the field of technology transfer[7] and the corresponding formulations in statutes of the individual centres, subsequent commercial use of the research results should also be enabled in general.

Quality as part of good scientific practice

The provision of research data for use in the long term requires a secured quality management system. This should cover the entire life cycle of the data and thus extend from the gathering of the data to workflows and processing methods, to the storage and securing of the data, and to its controlled deletion.

When gathering data, framework parameters (known as "metadata") should also be recorded alongside the description of the data recording procedure itself, which enable statements to be made regarding the quality of the data gathered. Which metadata needs to be gathered for this purpose depends in individual cases on the respective specific research activity/project. Information for quality assurance purposes can here be recorded both in the form of laboratory logs and also in the documentation of data-generating processes, for example.

[7] "We are helping to shape our future by connecting research and technology development with innovative application and precautionary perspectives."

However, in modern ("digital") scientific work, priority should be given to the explicit, machine-readable coding of this information, and this should be promoted – in theory, standardisation and efficiency-improving, unproblematic practice.

This machine readability becomes necessary in connection with automated data analysis at the latest, as does the explicit *and* machine-readable recording of error measures (whether as part of the data or of the metadata).

A comprehensive metadata record should also include information about the data formats used. As far as possible, open and free data formats should be used, since implementations can also be realised for these even after the original application used is no longer present. For subsequent use, particularly also with digital methods, it should be ensured that quality-assuring metadata is stored – as far as possible in digitalised form and accessible via algorithms.

Scientific acknowledgement

The generation of research data is a central and essential task in the research process, which benefits science and, indirectly, also society as a whole. Quality assured research data is therefore part of the scientific output of the centres and its generation should therefore be acknowledged.

Even so, the generation of research data is today often regarded as being of lesser importance than its analysis. This differentiation is often no longer appropriate, given the competences required for gathering and preparing data. An improved acknowledgement of the scientific work expressed in the gathering of research data therefore begins with the development of a new perspective on all work processes and the persons involved in each case, who can only produce excellent research results through cooperative effort.

Recording and evaluating approaches and activities in this collaborative process through established methods of performance measurement – particularly the publication of articles in highly-respected journals – can only be an interim step, since they perpetuate the deficits of established evaluation metrics and are in many cases unsuitable for recording the activity conducted.[8]

It is clear that the bibliometric indicators used in publishing are not sufficient in this context, and neither are the general, non-specific indicators such as the number of volume. For this reason, the Helmholtz Association is also pursuing the goal of developing suitable indicators with which scientific work completed within the scope of research data management can be made visible.

[8] See for example The Metric Tide, Report of the Independent Review of the Role of Metrics in Research Assessment and Management, July 2015. DOI: 10.13140/RG.2.1.4929.1363.
http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/The,Metric,Tide/2015_metric_tide.pdf

The issue of evaluation notwithstanding, quotable data publication not only guarantees traceable scientific acknowledgement, but also the reproducibility of the studies that are based on it. In addition, the Open Access to the research data is documented.

The following current examples (2016) prove the increasing acknowledgement of the scientific importance of research data.

- In its Proposal Guidelines since 2013, the American National Science Foundation no longer expects a list of articles (as long as possible), but 5 “products” each. These can be articles, data, software, or patents, which reflect the most important personal achievements or the most important results of the applicant for the progress of the community.[9]
- The development of the “Earth Systems Science Data” data journal is a clear example of how the scientific importance of data can also be made visible via the classic route of publications.[10]
- Increasingly, widely known, established journals are also demanding a “Data Availability Statement” for each article.[11] This demand will no doubt become a standard; it will depend on the publishers and assessors what type of “availability” is ultimately regarded as being acceptable.

Long-term availability

Research data should in most cases also be sustainably stored and secured, and if appropriate, published in a suitable, trustworthy research data infrastructure for long-term subsequent use, including after its first use.

For research data on which publications are based, a 10-year storage period should anyway be guaranteed on the basis of the “good scientific practice” rules. Furthermore, journals are increasingly demanding direct access to it.

When archiving and, if appropriate, rendering accessible research data, the implications that arise from the applicable law or contractual rights of third parties must be taken into account.

[9] See for example The Metric Tide, Report of the Independent Review of the Role of Metrics in Research Assessment and Management, July 2015. DOI: 10.13140/RG.2.1.4929.1363. http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/The,Metric,Tide/2015_metric_tide.pdf

[10] See The National Science Foundation, Proposal and Award Policies and Procedures Guide, Part I – Grant Proposal Guide (October 25, 2016), Chap. II, C.2.f(c). https://www.nsf.gov/pubs/policydocs/pappg17_1/nsf17_1.pdf

[11] Earth System Science Data. The Data Publishing Journal. http://www.earth-system-science-data.net.

Insofar as the necessary commitments have been fixed in the data management plan, nothing should in general stand in the way of implementing such long-term availability through a suitable data infrastructure. Subsequently determining rights – including moral rights – or usage scenarios is almost impossible, however.

- Under certain circumstances, multiple transfer of the data to different types of media is necessary in order to ensure long-term archiving. Parameters that can influence the quality of the data should be determined and recorded. This can be the case, for example, when compression methods are used that entail losses.
- Securing subsequent usability makes it necessary, particular with data that is stored in proprietary data formats, to guarantee access to software in order to read and process this data. The archiving of the relevant software – whether open or proprietary – should therefore be taken into account by a research data infrastructure when considering the prospect of long-term usage.
- Additionally, a possible use of formats and certainly also of metadata structures by new formats should also be taken into account. “Old” data must be adapted to this development if necessary.
- Without exception, a key component for archiving and the rendering accessible of the necessary metadata is the rights structure, which should also include a compilation of access rights throughout the entire life cycle of the data and beyond.
- In light of the huge increase in global data volumes, detailed specifications for the respective ends of the life cycles of the data are also important. It will not be financially possible to store each data item for an unspecified period of time (and for example beyond further technical and ontological developments).
- A portion of such a rights structure can also be a targeted deletion of data after a specified period. In this light, regulations should if necessary be set as to the form in which the data should be deleted following expiry of the life cycle, and how its deletion should be documented. (see also the chapter on legal issues)

Research data infrastructures

Sustainable research data management sets a wide range of technical and organisational standards, including standards applicable to the research data infrastructures. Their purpose is to guarantee effective work with the data, its long-term archiving, exchange with cooperation partners and the publication of data. Research data should be stored and archived in trustworthy research data infrastructures that are -internal to Helmholtz or external.[12]

Here, the term “research data infrastructures” always refers to the combination of hardware, software and competent staff. The competences built up to ensure secure operation, further development, quality

[12] http://www.re3data.org

assurance of the data and to create the research data infrastructures are at the same time a valuable reservoir for the qualification of staff at the centres, partners and communities.

Qualification

Due to the high level of importance and rapidly growing need for qualified staff for research data management and analysis in research and industry, the further development of training offers is currently an urgent necessity.

The centres – sometimes in cooperation with universities and/or businesses – provide their staff with a professional qualification and support programme that is oriented to the specific requirements of their particular discipline and infrastructure facilities.

Financing

The Helmholtz centres understand the need for ongoing additional work in order to provide research data management, and are meeting this need through structural and financial measures.[13] However, third-party funds supplement but do not form the basis of the financing of the research data management.

Legal issues

The realisation of Open Access to research data (see chapter “Open Access”) makes it necessary to also take a look at the relevant legal context.

Both the accessibility of research data in the spirit of Open Science and its commercial use mean that an inspection of the necessary right of disposal is required.[14] Here, labour law, the Employee Invention Act and the Basic Law (freedom of science) must be observed. It may be of relevance to ownership rights whether content to be archived is subject to copyright, whether it is a company or a commercial secret, or with which licence content should, in some cases, be exclusively forwarded or published. Other legal framework conditions may arise from fields as widely different as data protection or export controls.

The protection of personal data is naturally of particularly high importance with regard to (bio) medical data. The statutory regulations relating to this area are currently being revised by the new European General Data Protection Regulation. Clear, simple statements are therefore difficult or impossible to

[13] Advice for information infrastructures (3.V.2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, (“Performance from diversity, recommendations on structures, processes and financing for research data management in Germany”) Göttingen, Rat für Informationsinfrastrukturen, (German Council for Scientific Information Infrastructures) p. 63ff. www.rfii.de/?wpdmdl=1998

[14] Here, the term “right of disposal” is used, since for non-material goods, in contrast to material goods, the term “property” is not used. In the text, the term “property” is however used in some places, since in colloquial terms, it correctly expresses the fact that natural or legal entities should be named who may in some cases permit intended use.

make. However, the German Council for Scientific Information Infrastructures is currently working on a recommendation on this topic. The centres are advised to take this paper into account when designing their research data guidelines.

The items mentioned here are examples. They are intended to increase awareness among the centres of the necessity to check which issues are of relevance to them and how they should handle them. In general, the “Legal issues” appendix provides more detailed information about property ownership issues in the context of archiving and the accessibility of research topics. The Open Science working group endeavours to supplement this appendix with the aid of the Legal working group. Due to the complexity of the issues, however, it is currently not possible to say when this will be completed, however.

Appendix: Legal issues

The realisation of Open Access to research data (see chapter “Open Access”) makes it necessary to also take a look at the relevant legal context. Here, legal issues are explained with the goal of facilitating the broadest possible subsequent use of this data by third parties; this is also described as “legal interoperability”.[15]

Both the accessibility of research data in the spirit of Open Science and its commercial use mean that an inspection of the necessary right of disposal is required.[16] Here, labour law, the Employee Invention Act and the Basic Law (freedom of science) must be observed. It may be of relevance to ownership rights whether content to be archived is subject to copyright, whether it is a company or a commercial secret, or with which licence content should, in some cases, be exclusively forwarded or published. Other legal framework conditions may arise from fields as widely different as data protection or export controls.

It is of no relevance for issues relating to ownership rights whether or not the data is research data. Data – whether it is research data or not – is in the public domain, i.e. it is not subject to ownership rights. This does not apply when it is qualified as a work in the sense of copyright (see box “Threshold of originality and data”).

- The fact that data is in the public domain does not mean that no control can be exerted over its use. This control option is however not based on a right of ownership of the data, but on one or more contractual agreements. In the context of data gathering, these are typically labour or work contracts. In cases in which data is “purchased” from third parties – or to use more precise legal terms, the “purchaser” is given the data for use – the usage conditions including the level of a possible usage fee are regulated in a licence agreement. The licence agreement can for example determine how long the data may be stored and to whom it can be passed on.

Threshold of originality and data

According to German (!) law, a work that is protected by copyright is the product of a creative achievement with a minimum threshold of originality. The work must embody the creative achievement. Here, copyright protects not the content but the form of the work. Even if a great deal of creativity or creative energy has been used to generate data in the sense of factual or descriptive information, this is not embodied in the data as defined by copyright.

- Within the Helmholtz Association, these licence agreements are usually concluded between a centre and the data provider. The licence payer and therefore the person authorised to dispose within the scope of the contractual conditions is therefore the respective centre. If data is collected by staff at centres as part of their professional activity, as a rule, the right of disposal also lies with

[15] RDA-CODATA Legal Interoperability Interest Group (20.10.2016): Legal Interoperability of Research Data: Principles and Implementation Guidelines, https://zenodo.org/record/162241.

[16] Here, the term “right of disposal” is used, since for non-material goods, in contrast to material goods, the term “property” is not used. In the text, the term “property” is however used in some places, since in colloquial terms, it correctly expresses the fact that natural or legal entities should be named who may in some cases permit intended use.

the employer, in other words, the centre. This situation can become more complex when staff collect the data who are employed by more than one employer at the same time, or when data is collected as part of a project with third parties.

- The assertion that data – when it does not qualify as a work as defined by copyright – is in the public domain and requires an addendum. If data is stored within a database in accordance with the copyright laws (§§ 87a ff.), the database is subject to protection if it is inside the European Union. The database protection is an ancillary copyright for the investment in the creation of the database. It does not establish the right to ownership of its content. It binds the use of the database and the copying of key parts of its content to the approval of the database creator. The database creator is the natural or legal entity who has paid for the creation of the database. For this reason, within the Helmholtz Association, it can be assumed that a Helmholtz Centre is the database creator, either alone or together with project partners, and that it thus has the right to decide on the usage conditions of these databases.

From the above, it can be concluded that: during the planning stage of research projects, the information should already be compiled – ideally within data management plans – that is required for the clarification of ownership or disposal rights. Should the data be made accessible to third parties, information is supplemented with the forwarding or publication of the data as an item of metadata explaining who has the right of disposal over the data and, if appropriate, who the database creator is, and under which conditions the data or database may be used. For this purpose, international standard licences should be used, if available.

- In cases in which the data is subject to copyright, a subsequent usage option is particularly easily achieved through licensing with the CC BY licence. From the current Version 4.0 onwards, this licence also includes the simple transfer of any existing database rights. The licence permits commercial use of the licensed works. If this possibility is to be excluded, the licence is supplemented by the non-commercial amendment.

The CC 0 licence

The Creative Commons (CC 0) zero licence is the declaration that under certain circumstances, the owner of the rights waives existing copyright. Data that is in the public domain does not require such a licence. Even so, the use of the licence is recommended, since if the fact that the data is in the public domain is not in doubt, it does no damage and otherwise leads to the desired release.

- In cases in which data is intended to be published that is not subject to copyright protection – and this is usually the case! – it should be made the subject of the CC 0 (zero) licence. This licence declares that any existing rights to the content thus licensed are waived insofar as this is legally possible. Should this data be forwarded in the form of a protected database, the CC 0 licence also applies to it. In theory, the database can also be licensed separately. Licences such as the CC BY or the Open Database licence v1.0 (ODbL) are possible. A well-grounded recommendation for a certain licence for licensing the database is only possible when the individual circumstances are taken into consideration. In cases when data in the public domain is forwarded, the possibility of commercial usage can also be excluded by means of a contract.

- A possible disadvantage that arises in connection with the use of the CC 0 licence appears from the perspective of the scientists to be a waiver of the obligation to name the source. However, the removal of this obligation for copyright or contractual reasons does not release the user of the data from their obligations relating to good scientific practice. This requires that sources be named.
- Finally, it is possible that software should also be provided for the use of the data and/or the database. Software is subject to copyright and therefore requires a separate licence, e.g. in the form of the GNU GPL or Apache licences. Here, too, the individual circumstances must be taken into account before a licence can be recommended.[17]

If there are reasons not to use the licences explained above that are standardised for all cases, it should be ensured that an exclusive licence leaves the research data freely usable at least for scientific purposes.

A key point for the policy of every research data infrastructure is the decision as to whether data should be included about which the infrastructure cannot decide autonomously with regard to its further use. In this context, too, it must be ensured that it has been clarified and documented who has the right of disposal or the rights to the data, and on what legal basis (chain of rights) this situation is grounded.

A simpler issue is the question of ownership for operators of large-scale equipment, who only wish to retain and make publicly accessible data that is obtained during the use of their equipment, i.e. who do not want to take over external data. The “measured values” obtained through the use of the equipment are in the public domain. The operators of the large-scale equipment can determine in their usage regulations that the use of the equipment is dependent on agreement by the users to the free use by the operator of the data obtained. Should embargo periods apply, their length and the point in time when the period starts should be specified in the usage regulations. It is important that the agreement of the users to comply with the usage regulations is documented, and that this documentation is stored.

In order to minimise potential liability risks in the context of forwarding, the data should be labelled with a declaration which states that no guarantee shall be assumed that the data is correct (without warranty), and that the data is used at the user’s own risk. The precise wording of the declaration should be determined in coordination with the respective legal department of the centre. The centres should harmonise these declarations between each other as far as possible.

If the research data infrastructures of the centres accept personal data, compliance with the data protection regulations must be ensured. The centres as operators of such infrastructures are then responsible for compliance with the data protection regulations. As a result, standards may already arise for data suppliers at the data ingestion[18] stage, which should be explained in a policy of the research data infrastructure. Research data infrastructures that accept a broad spectrum of data for data ingestion, for which it is not always clear whether it also contains references to individuals should if

[17] The different types of licence are explained e.g. in Wikipedia: https://en.wikipedia.org/wiki/Apache_License

[18] The term “data ingestion” refers to the entire procedure for transferring data from the data supplier to the data infrastructure. This can simply entail uploading the data. However, it can also be a complex process that is connected to a conclusion of contract and/or curation of the data.

necessary obtain confirmation from the data suppliers that the imported data contains no reference to individuals.

HELMHOLTZ
Open Science