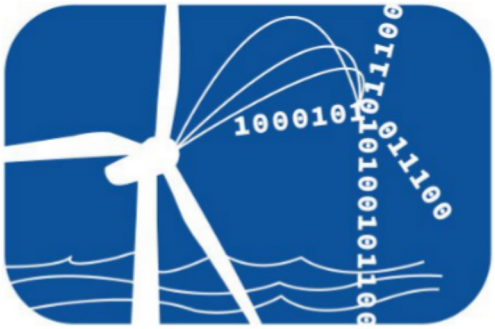

CRC 1463 Offshore-Megastructures

Project:	CRC 1463 Integrated Design and Operation Methodology for Offshore Megastructures
Title:	Guideline for Handling Research Data
Author:	WG Data Management
Last changed on:	17/01/2022
Adopted on:	
Version:	2

Table of contents

Preamble

Scientific research requires teamwork. Nevertheless, all researchers have their own way of working. To ensure effective collaboration, this document provides guidelines on how to handle research data across projects in the Collaborative Research Centre "Integrated Design and Operation Methodology for Offshore Megastructures" (SFB 1463). They ensure that partial results can be understood and processed equally well by all participants. This guideline defines minimum standards and is intended to give project participants security and orientation in handling of data.

This guideline applies from the start of the project to all project participants whose work requires scientific handling of research data. Research data involve all data that are collected and processed in the course of the scientific work process. They form the basis of scientific research results. This includes very different types of data, depending on the subject area. The CRC 1463 thus follows the understanding of research data of Leibniz Universität Hannover (LUH) and the German Research Foundation (DFG), which include data types such as measurement data, laboratory values, audio-visual information, texts, methodological test procedures, e. g. questionnaires, software and simulations [1, 2].

The CRC 1463 is committed to the importance of research data in the scientific knowledge process.

A reliable and responsible handling of research data is indispensable for the traceability of research and the dissemination of scientific knowledge. The research data management ensures the accessibility, re-usability, reproducibility and quality of research results. Therefore, the creation, processing, documentation, securing, storage and sustainable provision of research data should be carried out in accordance to recognized discipline-specific standards and be based on the FAIR Data Principles [3]. This ensures that data and metadata are findable, accessible, interoperable and re-usable.

In addition to this guideline, higher-level regulations apply. On the one hand, legal requirements such as the General Data Protection Regulation and copyright law must be complied with. On the other hand, the guidelines for ensuring good scientific practice of the DFG and LUH [4, 5] apply. Specific guidelines and directives on the handling of research data and on open access must also be complied with. These include the guidelines for handling research data [1, 2] applicable at LUH and explained by the DFG and the LUH Open Access Resolution [6].

If you have any questions regarding the implementation of these principles, please contact the Data Management Working Group (WG Data Management).

§1 Collaboration

With research data management, CRC 1463 pursues the common goal to make sure that the newly acquired research results are traceable and usable. Traceability is of great importance in interdisciplinary research projects. Especially the structure and accessibility of large amounts of data, which are generated through the interdisciplinary cooperation in the research area, shall be secured with the help of this guideline. At the same time, a common and uniform processing, storage and exchange of research data are indispensable. Especially when storing data, it should be taken care to well documentation and structuration so that colleagues who were not involved in creating the data can also derive an added value from the research results. Further details on the storage and naming of files can be found in §6. In this context, one of the most important aspects is the **joint** compliance with the data management goals. By implementing the goals of this guideline for interdisciplinary collaboration, successful use of the research data can be ensured for the project duration as well as for further funding periods and beyond. In the following paragraphs, the jointly pursued goals of CRC 1463 on data management will be explained and recorded.

§2 Plan for handling of research data

The planning and handling of research data is a main aspect of this guideline. It is important for interdisciplinary work that joint regulations and specifications are defined. Both technical and human resources must be aligned with the research project.

In general, research data are understood to be data that are created and/or subsequently used during the project (see §3). This includes, for example, measurement data, raw data, programming codes and calculation results. Classic research data always include metadata, which contains additional information about the available research data. It is described in more detail in §5.

For a structured resource planning in research data handling, a data management plan (DMP) should be established by all subprojects of CRC 1463. The DMP contains e. g. data format standards as well as delivery relations and delivery-time, which are to be defined specifically in subprojects. A generally applicable template to be used, will be developed by the Data Management WG and will be made available to all participants. The adaptation and creation of the DMP is in the responsibility of the individual subprojects and is to be adjusted specifically to them. With well-planned data management, that takes into account this guideline and the DMPs, scientific results will be traceable and measurable.

§3 Re-Use of Existing Data

The participating subprojects of the CRC 1463 are encouraged to check whether a re-use of existing data makes sense. Therefore, a research for relevant existing data should be carried out a so-called “secondary analysis”. When re-using existing data, the rules for correct citation have to be followed. When reproducing other research results such as articles, books, etc., it should be taken care to ensure that the correct details of author, publisher and year are given. The use of a literature management software (e. g. Citavi) can be useful when reusing large amounts of data. In any case, a correct and consistent citation method should be followed when publishing data.

The re-use of in the CRC generated data must also be subjected to a structured processing. Traceability is an essential aspect in the processing of research data, which can be ensured by additional metadata. The procedure for the selection of keepable and to be published data is described in §9.

§4 Compliance with Existing Standards and Documentation

In addition to this guideline, higher-level regulations and guidelines are valid without restriction and must be observed. These include in particular legal principles such as the European General Data Protection Regulation [7] or the copyright law [8]. In addition to the legal foundations, further guidelines and directives should be observed, which include the guidelines for ensuring good scientific practice of the DFG [2] and the LUH [1]. A main goal is to make research data accessible without restrictions. For this purpose, the LUH Open Access Resolution [1] and the Guidelines for Handling Research Data explained by the DFG should be followed.

When publishing research data, nondisclosure agreements (NDA) and similar agreements must be considered, especially but not only when dealing with sensitive data. Agreements made must be followed and considered when publishing research results, so that under certain circumstances only partial results can be published.

§5 Metadata

Metadata contains descriptions and details of the present research data. Metadata should facilitate the understanding and should support the further use and re-use of this research data. The term metadata comprises the information on the location, content, methodology, generation process, technology, documentation of required software, date/time, sources, identifications (e. g. DOI), etc. This detailed information should be complete for the present data and should be revised for adjusted data. The required information is thus depending on the type of data and could therefore change with different datatypes. When recording measurement data, for example, the measurement rate and the utilized measurement technology are to be considered in the metadata. Whereas the extensive metadata for software shall contain information on the operating system, version and similar aspects.

To clarify these above advantages, the metadata for the software DeSiO will be described as an example. DeSiO is a finite-element-software that will be further developed by the subproject Z01 and is utilized for the modelling with the Digital Twin Technology. Hence, the metadata of DeSiO should at least contain the following information:

- Details of the operating system on which DeSiO can be used
- Further system requirements (e. g. additional software)
- Version and the changes compared to former versions

- Input format (e. g. windIO)

In general, metadata should be captured and stored exactly as long as the corresponding generated research data [9]. It should be ensured that metadata is updated and made accessible according to the research data changes. As a format for the metadata, it is recommended to use the *.TXT or *.PDF file format. A catalogue for discipline-specific standards of metadata can be found in [10].

§6 Naming and Storage of Files

Research data should preferably be stored on the institute's own servers of the subprojects. The Seafile of the CRC 1463 can be used to store relevant data for the overall project such as protocols of meetings [11]. However, the storage capacity is limited to 40 GB and should therefore not be seen as an alternative to the institute server. A double insurance of data has to be ensured, e. g. through backups of the servers, cf. §7.

For the exchange of files, an exchange folder has been set up in Seafile, which is regularly cleared and should therefore not be used for a permanent storage. The path to it is:

Bibliotheken/sfb-megastrukturen/02_Projektphase/07_Austauschordner

If larger amounts of data is to be exchanged or sent, it is recommended to use the Download Ticket Service of LUIS Hannover [12]. Through the service all project participants can create a link to download the desired data, which can then be sent to the desired person (both LUH-internal and external persons).

GitLab is to be used for the exchange and storage of programming code [13]. As a part of the LUH project repository, the login information should be the same as that of project management via Seafile.

A convention for naming Seafile folders has already been established. When adding project folders in the first level (e.g. 07_exchange folder), the naming is based on the following principle: 00_folder name. Here, 00 should be replaced by a consecutive number that has not yet been assigned and the folder name should be given a meaningful title. The structure of the subordinate folder levels is not regulated with this guideline.

The naming of files follows a similar principle to that of folders. It was decided to comply with the following convention for file naming:

YYMMDD_AbbreviationEditorWithThreeLetters_NameofFile_v01.

Here the first six numbers stand for the creation date according to the format YYMMDD. Separated by an underscore, the three-letter abbreviation for the subproject, the editor or the working group/cluster is indicated. A corresponding table with abbreviations for working groups and clusters can be found in Table 1. Another underscore introduces the name of the file. The individual words are identified by upper and lower case letters. Each new word is introduced with a capital letter without a space. The last part of the file name indicates the version of the file. The abbreviation "v" stands for version and then a two digits number should be used here. Special characters (such as ä, ö, ß, etc.) are not to be used.

Table 1: Abbreviations for Working Groups and Clusters

Name of Working Group or Cluster	Abbreviations for file names
Research Data Management	RDM
Early Career Support	ECS
PhD Workshop	PHD
Design and Load Case Definition	DLC
Cluster Wind and Waves	CWW
Cluster Seabed integration	CSI
Cluster Support structure design	CSD
Cluster Rotor design	CRD
Cluster Control and Monitoring	CCM
Cluster coupled system	CCS

The name of this guideline is as follows:

210714_RDM_GuidelineForHandlingResearchData_v01

The file has a processing status of 14/07/2021, was created by the Research Data Management (RDM) WG and it is the first version.

§7 Protection Against Data Loss

Even before the actual data collection, it must be ensured that a sufficiently suitable infrastructure is available. The following questions are particularly important for this point:

- Is sufficient storage capacity available?
- Is an automated daily backup guaranteed?
- Is the bandwidth sufficient?
- Are the storage systems sufficiently secure?

For the daily backup, it is important to ensure that the backup copy is physically not located in the same building as the original, otherwise data can be lost as a result of unforeseen events

(such as fire). Furthermore, the bandwidth is important for the regular transfer of large amounts of data over an intranet or the Internet (e. g. for a distributed backup or for exchange with external partners). Data security is especially important for sensitive information. Aspects to protection against data abuse are described in §8 in detail.

In addition, LUH provides a LUIS services for storing and backing up data. For cross-project content and for sharing smaller amounts of data the project seafile is available as a project repository [11]. For long-term data retention, the LUH data archive should be used [14]. This data archive is intended for data that does not need to be published for legal and ethical reasons, but need to be kept due to good scientific practice. For data to be published where no suitable subject repository can be found, the LUH data repository should be used [15]. The data of the current project are to be stored daily on institute servers, either by working directly on the servers or by using software for automated synchronization of folders (e.g. PureSync).

§8 Protection Against Data Abuse

Protecting research data from unauthorized access is an important part of research data management. This paragraph explains which methods and tools are available. It discusses

- the handling of passwords,
- the management of rights,
- the encryption of files and data carriers and
- other organizational measures.

Dealing with passwords

There are numerous aspects to be considered when dealing with passwords. Firstly, it should be noted which aspects should be avoided by all means when dealing with passwords:

- Do not stick passwords to the screen
- Never send passwords in plain text by e-mail
- Do not use the same password for different accounts (at least not for sensitive data and infrastructure)

These basic recommendations prevent unauthorized persons from accessing passwords, access to sensitive data from being passed on by carelessly forwarded e-mails, and access to all files from a one-time "cracking" of the password.

The following tips should be followed for creating secure passwords:

- at least 20 characters
- difficult to guess (do not use birthdays, etc.)
- Passphrase (compilation of several words without recognizable meaning, i.e. no proverbs etc.) instead of a single word
- change spelling (e.g. "Tr!ck" instead of "Trick")
- combine several languages

Further tips can be found, for example, in the recommendations of the German Federal Office for Information Security (BSI) [16]. In addition, country-specific special characters such as ä, ö, ü, ß should not be used when creating passwords, as these characters are not available on keyboard layouts in other languages and the password can therefore not be entered.

Rights management

Some systems for data management offer the possibility to define access rights (e.g. Cloud-Seafile of LUH; currently not possible for Project-Seafile of LUH). Rights can be given to

individual files, folders or entire folder trees and data sets in a database. The simplest level is the assignment of

- combined read and write rights,
- read-only rights or
- no access rights.

In principle, it is advisable to ask IT specialists to reliably set up and manage permissions. This avoids incorrect configurations that can lead to data loss or security breaches. Rights management must be taken into account especially when external parties are to be given access to data. Since access rights cannot be defined for Project Seafile, and Cloud Seafile should not be used for project processing, access rights are particularly important for the institute servers, since many employees potentially have access to data here that may not be required.

Encryption of files and data carriers

Encryption of data may be required for legal and ethical reasons. This is the case, for example, with personal or patent-relevant data, but also with data that is subject to a NDA. The principle applies that the more sensitive the data, the more important end-to-end encryption is.

The following tips should be followed to ensure this principle:

- Encryption of USB sticks / external hard drives, as they are easily lost
- Encryption of hard disks integrated in computers (login password of the operating system is not sufficient, as the hard disk could be removed)
- sensitive data stored on institute servers or data centre servers are usually secure, as the servers are located in secured rooms
- no sending of sensitive folders or files via email or storage on commercial file hosters and cloud storages (e. g. Dropbox, Microsoft OneDrive etc.)

Software-based encryption should be used to encrypt laptops or external hard drives. This tends to be more secure than hardware-side encryption. LUH recommends the free open source program VeraCrypt, which works on most operating systems [17], but encryption tools adapted to the operating system (e.g. BitLocker for Windows) can also be used. If sending sensitive data by e-mail is unavoidable, the files should be encrypted in advance (e. g. with 7-Zip).

Most data abuse attempts are happening online. Nevertheless, physical abuse of access to data storage must also be considered. The following questions must be answered:

- On which data carriers do originals or copies of the data exist (e.g., e-mail attachments, paper printouts, external storage media or backup servers)?
- Where are the data carriers located?
- Who has legal physical access (colleagues, janitors, cleaning service, etc.)?
- How easy is it to gain illegal access to the data carriers (burglary)?

Storage locations for the data media should be selected whose security and access restrictions are appropriate to the sensitivity of the data.

Further organizational measures

In order to raise awareness of protection against data abuse all project participants should be trained.

Data media that are no longer needed or are defective should be professionally and physically destroyed to prevent subsequent recovery of the data. For this purpose, the data media can

be handed in, for example, to the LUH (room R318A) at Subject Area 12. The data carriers will be destroyed in an appropriate professional manner.

Regular checks to monitor the compliance with the rules will be done. The workflow and responsibilities are determined by each subproject itself. The defined boundary conditions are to be recorded in the DMP.

§9 Selection of Data to be Retained

In the spirit of good scientific practice, relevant data should be stored for the long term utilization and, if possible, made publicly available. Not all data must be retained. However, it must be consciously decided and justified which data shall be stored and which shall not. The following scenarios act as examples of data worth keeping.

If the data are the basis of a scientific publication, then these data must generally be retained for at least 10 years.

If subsequent use of the research data is likely, then it must also be archived.

Unique data, i.e. data that cannot be collected in the same form (e.g. weather records or opinion polls) should be retained.

If data cannot be reproduced at a reasonable cost, then the data should be retained. For example, simulations should always produce the same result if the software used, the algorithms used and the input data are the same. In this case, the data can be deleted. However, metadata information should be archived, in this example statements about the used software environment and input data. It should be ensured that, for example, calculation results are well documented in order to avoid storing all software versions.

High-quality data should be archived. Source data and end products/results should therefore always be stored. File versions that represent intermediate stages and are partially defective tend to be deleted. Metadata should be archived for the file versions to show that this data existed and why it was deleted.

In the case of very large files, the storage space required is sometimes too large to archive this data. First of all, try to increase the storage capacity or reduce the file size. Only in the next step the less important and older files should be deleted in favour of the more important and current files.

§10 Long-Term Archiving and Publication of Research Data

For the long-term archiving and publication of research data, the data must be prepared according to the FAIR principles. FAIR is an acronym for data that includes findable, accessible, interoperable, and re-usable.

Data preparation according to FAIR principles

Findability is intended to enable other researchers to find the desired data. Care should be taken to ensure that rich metadata are available in searchable public directories (e.g., search databases of data repositories) and that these data are retrievable via a unique identifier (e.g., DOI).

Accessibility allows other researchers to access the data created. If possible, online access via standard protocols (e.g. http(s)) should be considered. In addition, transparent access conditions should be established, justifying under which conditions access to the data is granted (if access is restricted).

Interoperability with other systems ensures that the generated data can be combined with other data and can be processed by machines. Accordingly, the data are systematically prepared and documented according to discipline-specific standards. Furthermore, the machine-readable data and metadata are to be archived in common and, if possible, open file formats and should be referenced to related data.

Re-usability enables other researchers to re-use the data for their own purposes. In addition to all the previous points, it is important to have a good documentation and a clear definition of re-usability conditions.

Archiving data

Proper archiving applies to data, metadata and other information. The basic rule is that

- Data is available in long-term readable file formats,
- data is structured according to discipline-specific or data-specific standards, and
- data should be as machine-readable as possible [10, 18].

For example, images should be saved in *.TIFF or *.TIF format, documents in *.TXT, *.PDF/A or *.XML, and audio files in *.WAV. Following §7, the following requirements are important for long-term archiving of data:

- suitable server rooms (fire protection and security requirements)
- redundant storage
- regular replacement of storage media (as a result of wear and tear of the data media)
- guaranteed bitstream preservation (error-free data preservation) for at least 10 years
- maintenance by IT specialists
- data access for authorized persons only

CDs, external hard drives or USB sticks are not suitable for long-term archiving!

Publication of high-quality data

If legally possible, data prepared according to FAIR principles should be published. This is not only in the interest of funders and scientific organizations, but also in the interest of the researcher, since relevant and high-quality data generates additional influence and attention.

Published data must be cited in the same way as text publications.

The data to be published must be made available via a so-called data repository (data centre). This repository stores and manages data and the associated metadata. The data can then be found via search criteria and can often also be downloaded. It is important to ensure that the repository is specialized in the subject or type of data in question. Suitable repositories can be found in [19]. The following aspects should be considered when selecting a suitable repository:

- Assignment of a Persistent Identifier (PI, e.g. DOI).
- Clearly regulated and binding data utilization
- Who is allowed to set up data?
- Which formats for data and metadata are accepted?
- What services does the repository offer (e.g. quality assurance)?
- What does it cost to set up data?

Ideally, these points should be clarified before the start of the project in order to be able to apply for the necessary funding and generate the right file formats. If no suitable data repository can be found, LUH offers a free data publication option via its own institutional data repository [20].

If the data publication is accompanied with a publication of a technical article describing the data evaluation and assessment, the metadata must refer to the other publication in each case.

References

[1] Leibniz Universität Hannover, *Richtlinie zum Umgang mit Forschungsdaten an der Leibniz Universität Hannover*, 2020.

[2] Deutsche Forschungsgemeinschaft, *Leitlinien zum Umgang mit Forschungsdaten*, 2015.

[3] FAIRplus, *FAIR Cookbook.* [Online]. Available: https://fairplus.github.io/the-fair-cookbook/content/home.html (accessed: Jun. 8 2021).

[4] Leibniz Universität Hannover, *Ordnung der Gottfried Wilhelm Leibniz Universität Hannover zur Sicherung guter wissenschaftlicher Praxis*, 2015.

[5] Deutsche Forschungsgemeinschaft, *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Denkschrift; Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft"*. Weinheim: Wiley-VCH, 2013.

[6] Leibniz Universität Hannover, *Open Access-Resolution der Leibniz Universität Hannover.* [Online]. Available: https://www.uni-hannover.de/de/universitaet/profil/ziele-strategien/open-access/open-access-resolution/ (accessed: Jun. 8 2021).

[7] Datenschutz-Grundverordnung (DSGVO), *Datenschutz-Grundverordnung: DSGVO als übersichtliche Seite.* [Online]. Available: https://dsgvo-gesetz.de/ (accessed: Jul. 10 2021).

[8] Bundesministerium der Justiz und für Verbraucherschutz, *UrhG.* [Online]. Available: https://www.gesetze-im-internet.de/urhg/ (accessed: Jul. 11 2021).

[9] University of Sydney, *Research Data Management Policy 2014.* [Online]. Available: https://www.sydney.edu.au/policies/showdoc.aspx?recnum=PDOC2013/337%20https://blog.imb.bau.tu-dresden.de (accessed: Jul. 10 2021).

[10] Research Data Alliance, *Metadata RDA | Metadata Directory.* [Online]. Available: https://rd-alliance.github.io/metadata-directory/standards/ (accessed: Jun. 22 2021).

[11] Leibniz Universität Hannover, *Projektverwaltung Seafile.* [Online]. Available: https://admin.projekt.uni-hannover.de/projekte/projects/view/project/sfb-megastrukturen (accessed: Jul. 11 2021).

[12] Leibniz Universität Hannover, *Download Ticket Service – Leibniz Universität IT Services – Leibniz Universität Hannover.* [Online]. Available: https://www.luis.uni-hannover.de/de/services/speichersysteme/dateiservice/download-ticket-service/ (accessed: Jul. 10 2021).

[13] Leibniz Universität Hannover, *GitLab – Leibniz Universität IT Services – Leibniz Universität Hannover.* [Online]. Available: https://www.luis.uni-hannover.de/de/services/speichersysteme/dateiservice/projektablage/teildienste-der-projektablage/gitlab/ (accessed: Jul. 2 2021).

[14] Leibniz Universität IT Services, *Archivierung.* [Online]. Available: https://www.luis.uni-hannover.de/de/services/speichersysteme/archivierung/ (accessed: Jun. 21 2021).

[15] Leibniz Universität IT Services, *Forschungsdaten-Repositorium.* [Online]. Available: https://www.luis.uni-hannover.de/de/services/speichersysteme/forschungsdaten-repositorium/ (accessed: Jun. 21 2021).

[16] Bundesamt für Sicherheit in der Informationstechnik, *Empfehlungen des Bundesamtes für Sicherheit in der Informationstechnik.* [Online]. Available: https://www.bsi.bund.de/DE/Home/home_node.html (accessed: Jun. 21 2021).

[17] VeraCrypt, *Download VeraCrypt.* [Online]. Available:

https://www.veracrypt.fr/en/Downloads.html (accessed: Jun. 21 2021).

[18] forschungsdaten.info, *Formate erhalten.* [Online]. Available: https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten/ (accessed: Jun. 22 2021).

[19] re3data, *re3data.org.* [Online]. Available: https://www.re3data.org/ (accessed: Jul. 11 2021).

[20] Leibniz Universität Hannover, *Research Data Repository.* [Online]. Available: https://data.uni-hannover.de/ (accessed: Jun. 22 2021).