Integrating Activities for Advanced Communities

D4.3 – INTERACT Field Guide to Data Repositories

Project No.730938– INTERACT

H2020-INFRAIA-2016-2017/H2020-INFRAIA-2016-1

Start date of project: 2016/10/01 Duration: 48 months
Due date of deliverable: 2018/09/30 Actual Submission date: 2018/11/10

Lead partner for deliverable: AWI, METNO
Author: Anna Irrgang, Esther Hemmens, Øystein Godøy

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

Table of Contents

Publishable Executive Summary

The Arctic is a key region for understanding global climate change. Thus, access to data collected at INTERACT research stations is crucial for the global research community. Providing access to collected data is one major obligation for projects funded by the European Union, like the Horizon 2020 project INTERACT. Therefore, data collected at INTERACT stations need to be made accessible for the public. INTERACT promotes free and open access to data in line with the European Open Research Data Pilot OpenAIRE. This field guide to data management and data repositories provides information about data management principles and their implementation in the workflow, preparation and maintenance of data for efficient discovery and reuse, as well as some guidance on how to choose an appropriate data repository for data preservation and publication in order to comply to the INTERACT open access principles.

1 Introduction

1.1 Background

Eighty-three INTERACT stations are distributed over 16 countries across the Arctic and the high Alpine regions in Europe. A wide range of scientific disciplines and research projects are represented, at which vast amounts of data are being collected and created, for example in the frame of short-term process studies, or long term monitoring activities. Since the Arctic is a key region for understanding global climate change, access to data collected at INTERACT research stations is crucial for the global research community. Providing access to collected and created data is one major obligation for projects funded by the European Union, like INTERACT as a Horizon 2020 project. Thus, data generated within the INTERACT network need to be made accessible for the public. INTERACT promotes free and open access to data in line with the European Open Research Data Pilot OpenAIRE[1] as is stated in the INTERACT Data Management Plan[2] and the INTERACT Data Policy[3]. However, complying with the Horizon 2020 requirement for open data is not unambitious. So far, the management of data collected at the INTERACT stations follows no unified strategy. A survey carried out by the INTERACT Data Forum in 2017 revealed a very heterogeneous distribution of knowledge about data management among the INTERACT stations. At some stations, a data management plan is already implemented and the data management follows well-coordinated workflows. At other stations, some guidance is still needed. Providing guidance to establish a unified data management strategy is the task of the INTERACT work package 4 – the Data Forum.

1.2 Purpose

The purpose of this field guide is to increase knowledge on the topic of data management and to raise awareness for its importance. It intends to help station managers, as well as individual researchers to embed data management into their scientific workflow, to provide open access to the collected data and secure their long-term storage in a trusted data repository. This is the only way to avoid redundancy of activities like data collection or data analysis and loss of valuable research data. Further, it aims to encourage station managers to establish and implement standardized data management practices at their stations. The overarching aim is to improve data management practices across all INTERACT stations at all levels in order to ensure long-term accessibility, usability and preservation of collected data and thus also increase visibility of INTERACT stations as trustworthy data providers.

1.3 Scope

The audience of this field guide are researchers and further people who are collecting and creating data by using INTERACT field stations, station managers, as well as third parties contributing to INTERACT or using data supplied through INTERACT.

[1] OpenAIRE: https://www.openaire.eu/item/open-research-data-pilot-in-h2020
[2] INTERACT data management plan – deliverable 4.1: https://eu-interact.org/app/uploads/2017/11/D4._1_v1.pdf
[3] INERACT data policy – deliverable 4.4. No link assigned, yet.

2 Data management principles

Data management principles help researchers to keep track of their data and to handle and store them in a sustainable manner throughout as well as after the project. Thereby, data accessibility and reusability by humans and machines is ensured. In 2016, the so called FAIR guiding principles were published by Wilkinson et al.[4] and quickly established as the best practice guidelines for management of scholarly data.

2.1 FAIR Guiding Principles

The FAIR Guiding Principles are a set of 15 concise and measurable principles. They are summarized in four categories, being: **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR) (Wilkinson et al., 2016). The European Commission released a document with guidelines on FAIR Data Management in Horizon 2020 projects[5]. The GoFAIR initiative provides a detailed description of every principle[6] at their webpage. FAIR data are:

Findable:
F1. (meta)data are assigned a globally unique and eternally persistent identifier
F2. data are described with rich metadata
F3. (meta)data are registered or indexed in a searchable resource
F4. metadata specify the data identifier

Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1. the protocol is open, free, and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

Reusable:
R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards. (Wilkinson et al., 2016)

These principles are making data management measurable. They intend to enhance the findability of data for humans and machines and to increase data reuse. The European Research Council

[4] Wilkinson et al. 2016: The FAIR Guiding Principles for scientific data management and stewardship: https://www.nature.com/articles/sdata201618.
[5] European Commission guidelines on FAIR data managment in Horizon 2020: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[6] GoFAIR initiative – FAIR principles: https://www.go-fair.org/fair-principles/

published guidelines on how to implement open access to scientific publications and research data in the framework of Horizon 2020 projects[7].

2.2 Implementation of FAIR guiding principles in the workflow

In order to implement the FAIR guiding principles in the workflow, it is recommended to create a data management plan prior to data collection at the beginning of the research project. The main purpose of a data management plan is to make research data created within a project FAIR. It should involve a documentation of how data will be handled during and after the end of a research project, including information on the methodology and standards which will be applied and how data will be curated and preserved in the long-term. Its extent and content is varying depending on the funding agency, the size of the project and the amount and nature of the collected data. Guidelines on how to create a data management plan can be found for example on the website of the Data Curation Centre (DCC) [8]. The DCC also provides an online data management planning tool[9] which allows multiple project partners to work on the data management plan together.
The implementation of the FAIR guiding principles is a process that starts before data collection and goes beyond their storage. The simplified life cycle of data is displayed in Figure 1. Its central message is that data life is a cycle and handled sustainably, can lead to data re-cycling thus raising their use and value by exploiting them maximally.

Figure 1: Life cycle of data. Own creation (CC-NC-SA).

Planning: Already in the planning phase of the research project, it is helpful to think through the data life cycle, starting from how to collect them and ending on how to preserve them. As mentioned above, a data management plan can help to organize project data, keep track of them during the course of the project, as well as regulate data sharing and preservation during and after the project.

[7] European Research Council – Guidelines on Implementation of Open Access to Scientific Publications and Research Data: http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf
[8] Digital Curation Centre – management plans: http://www.dcc.ac.uk/resources/data-management-plans
[9] Digital Curation Centre – online tool for DMP creation: https://dmponline.dcc.ac.uk/

Collection: Prior to data collection, the creation of a data sampling plan may help to improve data collection by ensuring that data will be collected following standards set in the respective field or research. Usage of best practices for data sampling may reduce additional work to standardize data afterwards. During data collection, it is important to do a thorough metadata documentation. Metadata are data describing data (see Chapter 3). Metadata standards and schemas differ, depending on the type of data generated. The usage of controlled vocabulary, where applicable, is also recommended in order to raise the value of the data entry. The usage of open data formats, metadata standards and controlled vocabulary ensures the findability and interoperability of data.
Processing: Data processing implies its modification. Processing steps need to be included in the metadata. The documentation of data evolution is crucial for the quality of the data product, as well as for the reproducibility of the results. A data versioning and a regular data backup during processing counteract data loss. For backing up data, in addition to a personal device (computer, external hard drive etc.) an external storage system should be chosen, like for example a university server, or a cloud storage system. A good data backup plan also implies that data are stored in two places which are physically apart, like for example in the office and in the institutional server room.
Preservation: The preservation of research data is a decisive step in the data life cycle. The best practice is to store any relevant project data together with its metadata in a trusted data repository. Criteria of trusted data repositories are described in Chapter 4 of this document.
Publication: Data publication in a trusted data repository is a necessary step to ensure data accessibility. Either the data and its corresponding metadata, or just the metadata can be released for publication. INTERACT promotes free and open access to data in line with the European Open Research Data Pilot OpenAIRE. Embargos and licensing provide control over data access and usage. If the data set contains personal sensitive data, like for example names or birth dates, the open software Amnesia[10] from OpenAIRE can be used in order to anonymize these data and thus make them suitable for publication. In some cases, access restrictions may apply to data, for example when data release would affect intellectual property rights. A listing of cases in which the INTERACT open data principle do not apply, can be found in the INTERACT data policy.
Reuse: For the reuse of data, they need to be made accessible in an interoperable format and documented in a self-explanatory way. Rich and standardised metadata and open data formats increases the value of data and the likelihood for its reuse.

2.3 Interoperability of INTERACT station data

In Wilkinson et al., (2016) Interoperability is described as:

"The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort."(Wilkinson et al., 2016: 2)

In the context of data management at INTERACT stations this means, that data should be collected and processed in a way that they are usable beyond the framework of the respective research project. This ensures the integration of INTERACT station derived data into other research workflows without much effort with a mutual benefit for data providers and data consumers. Following the principles below, will help to increase data interoperability:

[10] Amnesia anonymisation software: https://amnesia.openaire.eu/

1. Follow established “best practices” for data collection
2. Use standardized data sheets (if applicable)
3. Use standardized metadata formats
4. Store and publish data open data formats (data formats which are non-proprietary, meaning they are not controlled by a company or organization)
5. Publish data and/or metadata in a trusted data repository

2.4 Parties involved in data handling

Within the data life cycle, different parties are handling them. Each of these parties has different obligations and requirements towards the data, as well as benefits arising from data publishing.

2.4.1 Data Providers

All researchers who collect data at an INTERACT station, or create data based on data collected at INTERACT stations are potential data providers. Each data provider needs to document the collected and/or created data with metadata and store them in a trusted data repository in order to make the data findable, accessible and reusable for potential data consumers. Some benefits of publishing data in a data repository are, that the data provider and the data are permanently connected through the repository which makes the data providers ownership unquestionable and that the publication of research data increases the trustfulness of the research findings and enhances the chance for new collaborations.[11]

2.4.2 Data Repositories

Data repositories are web based services for the storage of data. Data repositories are important actors in terms of long-term preservation, data discovery and data reuse. Data are accessed and stored using software protocols that are suitable for both, human operators as well as machines. Once submitted to a data repository, data security, accessibility and curation should be ensured by the repository. Trusted data repositories commit themselves to the FAIR guiding principles. A set of criteria for data repositories and some recommended data repositories is listed in Chapter 4 of this document.

2.4.3 Data Consumers/ Re-users

Data consumers are the end users of the data and thus cover a wide spectrum of potential users, from researchers to the general interested public. What unifies them is their need for online data access. Therefore, data needs to be findable, readable and accessible.
Further data consumers need to understand the content, structure and purpose of the data, for which it needs to be provided in a self-explanatory way. The use of data sets which are not properly documented and thus cannot be interpreted and understood by other users is very limited. The most important duty for data consumers is to respect the creative commons attribution under which the data are released in order to ensure proper citation. Data re-users are also highly encouraged to make their newly created data FAIR.

[11] Data management guidelines for researchers: https://www.futuretdm.eu/practitioner-guidelines/data-management-guidelines-researchers/#1497191296318-578ddb4b-1783

3 Preparing and maintaining data for efficient discovery and reuse

3.1 Purpose

Data are a valuable resource - their collection, preparation and analysis require time, as well as human and monetary resources. Without proper documentation, structure, format, and maintenance their reusability is limited which reduces their value. If it comes to field data, the informational loss can be severe, since most observations cannot be repeated under the same conditions.

3.2 Metadata

Metadata provide the context for data and contain descriptive information about the dataset (e.g., author, location, time, methods used, etc.) so that the data can be inventoried, discovered and re-used. To ensure findability and set some preconditions for the reusability of the data, it is important to use established data and metadata standards. In general, for INTERACT station derived data this metadata standard should be Dublin Core, or ISO 19115 compliant. ISO 19115 "defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services"[12]. A list of used metadata formats depending on the scientific discipline can be found at the Research Data Alliance Metadata Directory website[13]. Metadata documentation should be a routine step in the workflow of data collection/creation and data processing. Metadata are acquired by the data repository during the process of data submission. In addition to making use of a metadata standard, the usage of controlled vocabulary may contribute to make data more findable and better understandable. Controlled vocabulary provides a consistent way to index data (e.g., ontologies, taxonomies).

3.2.1 Discovery metadata

Discovery metadata are metadata which enable machines to discover data. This is needed for example for finding data using a search engine online. Trusted data repositories provide the submitted metadata in a format that is machine readable and thus discoverable. Making data findable is the first principle of the FAIR data guiding principles and sets the precondition for making data accessible.

3.3 Data curation

Data curation is the management of data throughout their lifecycle with the goal of maintaining data for efficient reuse. This implies keeping them discoverable, retrievable and reusable while keeping their quality and thus add to their value. It "provides the *methodological* and *technological* data management support to address *data quality issues* maximizing the usability of the data" (Freitas and Curry, 2016: 87). Data curation is an active and ongoing process which requires the investment of time and resources beyond the scope of the respective research project for which the data were collected. It is the crucial process which keeps data Reusable. Without ensuring

[12] ISO 19115-1:2014 – Geographic information metadata: https://www.iso.org/standard/53798.html
[13] Research Data Alliance Metadata Directory: http://rd-alliance.github.io/metadata-directory/standards/

long-term reusability of data, the work done at the INTERACT stations may get unusable in the greater scientific context.

4 Criteria for trusted data repositories

Data publications ensure that data remains accessible and usable while providing acknowledgement to their creators. Thus, it is always a good decision to publish data, however, it is not irrelevant where they are published. There are large differences between data repositories, if it comes for example to supported data formats, data handling, data curation and long-term storage. A set of requirements was established which need to be fulfilled by a data repository in order to be recommended for usage for INTERACT station derived data.
Some of the most important are:

- Support of the **FAIR guiding principles**. If a data repository complies to the FAIR guiding principles or not is written in the data repository's data policy. Since the FAIR guiding principles got introduced to the data science community in 2016, some data repositories are still in the process of making their platform FAIR.
- Providence of a **persistent identifier**, to make data identifiable, assignable and citable. Personal identifiers are important to fulfil the *findability* part of the FAIR data principles. A persistent identifier is linked to the source page of a digital object (the data), even if the source is moved to another place. The most common used persistent identifier for research products (scientific publications, scholarly data) is the Digital Object identifier (DOI).
- Providence of data **licensing**, to specify the conditions of data reuse and to protect the author's intellectual property. Licensing is important to fulfil the *reusability* part of the FAIR guiding principles. Depending on the type of information, different ways exist to license data. Many trusted data repositories make use of the creative common licences[14] or of licenses which comply to them, like for example Open Data Commons[15]. Data licenses provide different degrees of data reuse restrictions. Information about the licensing type used by a data repository are provided in the data repository's data policy.
- Support of relevant **open data formats** and **metadata standards**, to ensure data usage is not impeded by proprietary data formats. The fulfilment of these criteria supports the *interoperability* and *reusability* part of the FAIR guiding principles. Further, the usage of relevant metadata standards supports the *findability* requirement.
- Possession of a **long-term mandate**, to counteract data loss over time by providing long-term storage. This ensures long-term *accessibility* and thus also *reusability* of data. A long-term mandate can only be guaranteed with sustainable funding, which is often not directly visible. The number of users and stored data sets may serve as an indicator for the stability and persistence of a data repository.

The search for a suitable data repository usually starts at the data repository of the own institution or research community, but there are also national, international, generalist and discipline-specific data repositories which are worth exploring. Institutional repositories usually provide sufficient structures to publish metadata related to published research papers and deposit data for storage (and thus backup). If the institutional data repository fulfils the abovementioned requirements, no

[14] Creative Commons: https://creativecommons.org/
[15] Open Data Commons: https://opendatacommons.org/

further action is needed. If not, metadata and data preservation and publication in a different data repository is strongly recommended. An advantage of discipline-specific repositories is that specific data formats and well established metadata standards are being used which allow a more specific data documentation which suits the needs of the respective research community. A disadvantage of these repositories is, that access might be restricted to special institutions or research branches. General data repositories, also called all catch repositories, support more general metadata standards and are open to more different types of information. A standard for general research data is the DataCite metadata standard, which is for example being used for discovery metadata entries of repositories on the data repository search platform re3data[16]. Zenodo is an example for an all catch repository. An advantage of such general repositories is that they usually do not have any access restrictions, a disadvantage might arise from limited to no support of discipline specific data formats and metadata standards.

4.1 Examples for thrusted data repositories

Following the abovementioned recommendations, a selection of thrusted data repositories is listed below, which are recommended for the storage of INTERACT station-derived data. These repositories were synthetized, using the data repository search platform re3data. For data which cannot be clearly assigned to one of the supported research disciplines, Zenodo may be an alternative worth exploring. Zenodo is an interdisciplinary open data repository which was created by OpenAIRE and CERN.

[16] Re3data – data repository search platform: https://www.re3data.org

Table 1: List of trusted data repositories recommended for preservation and publication of INTERACT station derived data. Data repositories appear in alphabetical order.

Name and Link	Research Subjects	Supported content types
4TU.Centre for Research Data http://researchdata.4tu.nl/en/home	Agriculture Analytical Chemistry Basic Biological and Medical Research Bioinformatics and Theoretical Biology Biology Biophysics Chemistry Computer Science Construction Electrical and System Engineering Engineering Engineering and Architecture Forestry Geochemistry Geography Geophysics and Geodesy Geosciences Horticulture and Veterinary Medicine Landscape Planning Life Sciences Logistics Materials Science Mathematics Method Development (Chemistry) Mineralogy and Crystallography Physics Soil Sciences Spatial Planning Systems Engineering Traffic and Transport Systems Transportation and Infrastructure Planning Urbanism Water Research	Archived data Audio-visual data formats Images Plain text Raw data Scientific and statistical data Standard office documents Structured text
DRYAD http://datadryad.org	Agriculture Anthropology Applied Microbiology Basic Biological and Medical Research Biochemistry and Animal Physiology Bioinformatics and Theoretical Biology Biology Ecosystem Evolution Analysis Forestry General Genetics Geology and Palaeontology Geosciences (including Geography) Horticulture and Veterinary Medicine Humanities and Social Sciences Life Sciences	Formats Plain text text Scientific and statistical data Software applications Structured Source code Standard office documents

	Medicine Microbial Ecology Microbiology Natural Sciences Plant Ecology Plant Sciences Social and Behavioural Sciences Virology and Immunology Zoology	
Figshare http://figshare.com	Engineering Sciences Humanities and Social Sciences Life Sciences Natural Sciences	Archived data Audio-visual data Images Plain text Raw data Scientific and statistical data Source code Standard office documents Structured graphics
Marine Geoscience Data System http://www.marine-geo.org/index.php	Atmospheric Science Geography Geology and Palaeontology Geosciences Natural Sciences Oceanography Water Research	Archived data Databases Formats Images Plain text Raw data Scientific and statistical data Standard office documents Structured graphics Audio-visual data Structured text
NSF Arctic Data Center https://arcticdata.io	Agriculture Atmospheric Science Basic Forest Research Biology Chemistry Forestry Geochemistry Geography Geosciences Horticulture and Veterinary Medicine Humanities and Social Sciences Life Sciences Mineralogy and Crystallography Natural Sciences Oceanography Physics Soil Sciences Water Research Zoology Plant Sciences	Archived data Audio-visual data Images Network based data databases Plain text Raw data Scientific and statistical data Software applications Source code Standard office documents Structured graphics Structured text
PANGAEA https://www.pangaea.de	Atmospheric Science and Oceanography Biology Geochemistry Geology and Palaeontology	Archived data Audiovisual data Images Plain text

	Geophysics Geosciences (including Geography) Mineralogy and Crystallography Natural Sciences Oceanography	Standard office documents
Polar Data Catalogue https://www.polardata.ca/	Atmospheric Science and Oceanography Geography Geosciences (including Geography) Health Services Research Humanities and Social Sciences Medicine Medicine Life Sciences Natural Sciences Oceanography Public Health Social and Behavioural Sciences Social Medicine Social Sciences	Images Scientific and statistical data Structured text
Scholars Portal Dataverse https://dataverse.scholarsportal.info/dataverse/sp	Engineering Sciences Humanities and Social Sciences Life Sciences Natural Sciences	Databases Formats Images Raw data Scientific and statistical data Standard office documents Structured text
Zenodo https://zenodo.org	Engineering Sciences Humanities and Social Sciences Life Sciences Natural Sciences	Archived data Audio-visual data Images Network based data Plain text Raw data Structured text Scientific and statistical data Source code Standard office documents Structured graphics

5 Tools, services, references

[1] OpenAIRE: The open research data pilot for H2020: https://www.openaire.eu/item/open-research-data-pilot-in-h2020

[2] Godøy, Ø., Radosavljevic, B., Biskaborn, B., 2017: INTERACT data management plan – deliverable 4.1: https://eu-interact.org/app/uploads/2017/11/D4._1_v1.pdf

[3] Godøy, Ø., 2018: INERACT data policy, deliverable 4.4, No link assigned, yet.

[4] Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., et al., 2016: The FAIR guiding principles for scientific data management and stewardship. Nature Scientific Data, 3: 160018, DOI: 10.1038/sdata.2016.18, https://www.nature.com/articles/sdata201618

[5] European Commission, Directorate, 2016: Guidelines on FAIR data management in Horizon 2020, Version 3.0: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[6] GoFAIR initiative: An Initiative to promote FAIR data: https://www.go-fair.org/fair-principles/

[7] European Research Council, 21 April 2017: Guidelines on Implementation of Open Access to Scientific Publications and Research Data, Version 1.1: http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf

[8] Digital Curation Centre (DCC): provides services and informative literature for research organisation and digital curation: http://www.dcc.ac.uk/resources/data-management-plans

[9] DCC: Online tool for data management plan creation: https://dmponline.dcc.ac.uk/

[10] Information Management Systems Institute (IMSI), Amnesia, 2016: Free software for anonymisation of person related research data: https://amnesia.openaire.eu/

[11] The Future of TDM: An initiative to improve the uptake of Text and Data Mining (TDM) in the European Union: Data management guidelines for researchers: https://www.futuretdm.eu/practitioner-guidelines/data-management-guidelines-researchers/#1497191296318-578ddb4b-1783

[12] International Organization for Standardization (ISO): ISO 19115-1:2014, a metadata standard for geographic information metadata: https://www.iso.org/standard/53798.html

[13] Research Data Alliance (RDA): Directory of discipline specific metadata standards: http://rd-alliance.github.io/metadata-directory/standards/

[14] Creative Commons (CC) licensing: https://creativecommons.org/

[15] Open Data Commons: Tools for providing and using open data: https://opendatacommons.org/

[16] Re3data – data repository search platform: https://www.re3data.org

Freitas, A. and Curry, E., 2016: Big Data Curation *in* Cavanillas, J.M., et al. (Eds): New Horizons for Data-Driven Economy, Chapter 6, p. 87-118. DOI 10.1007/978-3-319-21569-3_6.

All links were accessed for the last time at November, 9th, 2018.