D1.2 Data Management Plan DESIR DARIAH ERIC Sustainability Refined

Marco Raciti, Inês Queiroz, Carsten Thiel, Robert Jäschke

D1.2 Data Management Plan

DESIR

DARIAH ERIC Sustainability Refined

INFRADEV-03-2016-2017 - Individual support to ESFRI and other world-class research infrastructures

Grant Agreement no.: 731081

Date: 30-06-2017
Version: 1.0

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081.

Grant Agreement no.:	731081
Programme:	Horizon 2020
Project acronym:	DESIR
Project full title:	DARIAH-ERIC Sustainability Refined
Partners:	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS UNIVERSITEIT GENT UNIWERSYTET WARSZAWSKI FACULDADE DE CIENCIAS SOCIAIS E HUMANAS DA UNIVERSIDADE NOVA DE LISBOA CENTAR ZA DIGITALNE HUMANISTICKE NAUKE GOTTFRIED WILHELM LEIBNIZ UNIVERSITAET HANNOVER INSTITUT NATIONAL DE RECHERCHE ENINFORMATIQUE ET AUTOMATIQUE KING'S COLLEGE LONDON UNIVERSITY OF GLASGOW KNIHOVNA AV CR V. V. I. HELSINGIN YLIOPISTO SIB INSTITUT SUISSE DE BIOINFORMATIQUE UNIVERSIDAD NACIONAL DE EDUCACION A DISTANCIA UNIVERSITY OF HAIFA
Topic:	INFRADEV-03-2016-2017
Project Start Date:	01-01-2017

Project Duration:	36 months
Title of the document:	Data Management Plan
Work Package title:	Dissemination and Innovation
Estimated delivery date:	30-06-2017
Lead Beneficiary:	DARIAH
Author(s):	Marco Raciti [marco.raciti@dariah.eu] Inês Queiroz [qines@fcsh.unl.pt] Carsten Thiel [thiel@sub.uni-goettingen.de] Robert Jäschke [r.jaschke@sheffield.ac.uk]
Quality Assessor(s):	Marco Raciti [marco.raciti@dariah.eu] Stefan Buddenbohm [buddenbohm@sub.uni-goettingen.de]
Keywords:	Data Management Plan, FAIR data, Open Access, OPRD

Revision History

Version	Date	Author	Beneficiary	Description
0.1	15-06-17	Marco Raciti	DARIAH	First draft
0.2	20-06-17	Carsten Thiel, Inês Queiroz, Robert Jäschke	UGOE, FCHS	Datasets added
1.0	30-06-17	Marco Raciti	DARIAH	Final version

Table of Content

Executive Summary

This document is the first version of the Data Management Plan (DMP) for data collected and created by DESIR. It describes the datasets generated during the course of the project, how the data will be produced and analysed. It details also how the data generated will be shared, disseminated and preserved.

	Nature of the deliverable	
	R	Document, report
	DEM	Demonstrator, pilot, prototype
	DEC	Websites, patent fillings, videos, etc.
	OTHER	
✓	ORDP	Open Research Data Pilot
	Dissemination level	
✓	P	Public
	CO	Confidential only for members of the consortium (including the Commission Services)
	EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Disclaimer

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

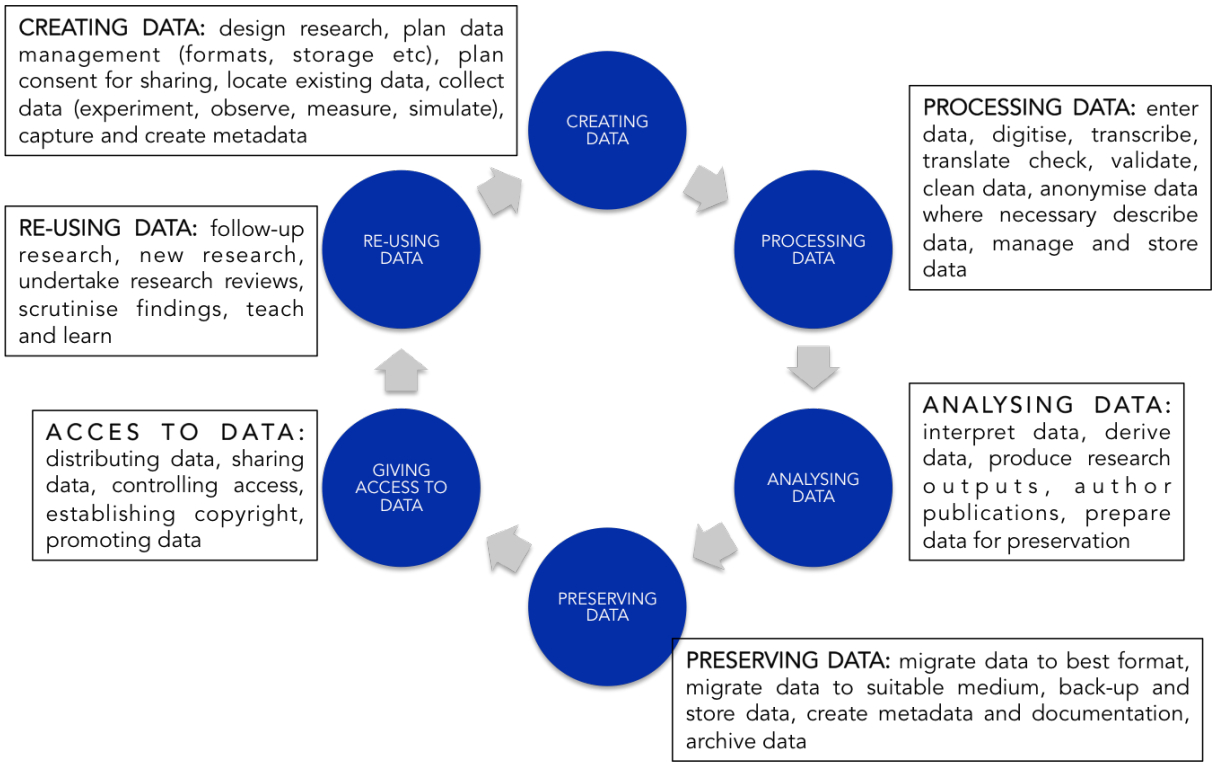
Introduction

The DESIR project sets out to strengthen the sustainability of DARIAH and firmly establish it as a long-term leader and partner within arts and humanities communities. By DESIR's definition, sustainability is an evolving 6-dimensional process, divided into the following challenges:

1. Dissemination: DESIR will organise a series dissemination events, including workshops in the US and Australia, to promote DARIAH tools and services and initiative collaborations.
2. Growth: DESIR sets out to prepare the ground for establishing DARIAH membership in six new countries: the UK, Finland, Spain, Switzerland, Czech Republic and Israel.
3. Technology: DESIR will widen the DARIAH research infrastructure in three areas, vital for DARIAH's long-term sustainability: entity-based search, scholarly content management, visualization and text analytic services.
4. Robustness: DESIR will make DARIAH's organizational structure and governance fit for the future and develop a detailed business plan and marketing strategy.
5. Trust: DESIR will measure the acceptance of DARIAH, especially in new communities, and define mechanisms to support trust and confidence in DARIAH.
6. Education: Through training and teaching DESIR will promote the use of DARIAH tools and services.

Funded under the Work Programme part European Research infrastructures (including e-Infrastructures), DESIR participates to the Open Research Data Pilot (ORDP). As such, this document introduces the first version of the project Data Management Plan (DMP). The DMP describes the data management life cycle for the data to be collected, processed and/or generated by the DESIR Consortium. It includes information on how the research data will be handled during and after the end of the project, what data will be collected or generated, which methodology and standards will be applied, how the data will be disseminated and shared, how the data will be curated and shared during and after the end of the project.

The DMP is to be considered as a living document: the information contained therein will be updated following the implementation of the project and when significant changes occur. Moreover, in a second phase, the Consortium intends to deliver a version which will contain recommendations to the community and specifically tailored to Digital Humanities. Research data generated and processed in DESIR will be FAIR (findable, accessible, interoperable and reusable).



Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

1. What kind of data is considered in the DMP

According to the *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020* (Version 3.0, 26 July 2016), research data "*refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form.*"

DESIR focuses on the sustainability of the DARIAH ERIC and will develop Actions consisting primarily of accompanying measures such as standardisation, dissemination, awareness-raising and communication, networking, coordination, support services, policy dialogues and mutual learning exercises. As such, we expect that limited research data will be generated during the course of the project.

2. Structure of the template

Each dataset will cover issues identified in the template below. The data will follow best practises defined in the *Guidelines on FAIR Data Management in Horizon 2020* to make the DESIR research data findable, accessible, interoperable and re-usable[1].

Data Summary	• State the purpose of the data collection/generation • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any) • Specify the origin of the data • State the expected size of the data (if known) • Outline the data utility: to whom will it be useful
2. FAIR Data 2.1. Making data findable, including provisions for metadata	• Outline the discoverability of data (metadata provision) • Outline the identifiability of data and refer to standard identification mechanism. • Do you make use of persistent and unique identifiers such as Digital Object Identifiers? • Outline naming conventions used • Outline the approach towards search keyword • Outline the approach for clear versioning • Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
2.2 Making data openly accessible	• Specify which data will be made openly available? If some data is kept closed provide rationale for doing so • Specify how the data will be made available • Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? • Specify where the data and associated metadata, documentation and code are deposited • Specify how access will be provided in case there are any restrictions

[1] EUROPEAN COMMISSION, Guidelines on FAIR Data Management in Horizon 2020, Version 3.0 (2016), p. 4

2.3. Making data interoperable	• Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. • Specify whether you will be using standard vocabulary for all data types present in your dataset, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
2.4. Increase data re-use (through clarifying licences)	• Specify how the data will be licenced to permit the widest reuse possible • Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed • Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why • Describe data quality assurance processes • Specify the length of time for which the data will remain re-usable
3. Allocation of resources	• Estimate the costs for making your data FAIR. Describe how you intend to cover these costs • Clearly identify responsibilities for data management in your project • Describe costs and potential value of long term preservation
4. Data security	• Address data recovery as well as secure storage and transfer of sensitive data
5. Ethical aspects	• To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former
6. Other	• Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

3. Datasets description

The DESIR Consortium identified the datasets that will be produced during the different phases of the project. The list is, however, to be considered indicative and datasets may be adapted, added or removed following the evolution of the project.

3.1 Datasets

1) Source code of software developed in WP4
2) Survey conducted in WP6

Dataset 1: Source code of software developed in WP4	
Data Summary	Computer source code generated as part of the programming activities in WP4, through development activities by the project partners contributing to WP4 and through the Code Sprint planned for Summer 2018. The efforts will follow the recommendations to encourage best practices in research software: https://softdev4research.github.io/recommendations/
2. FAIR Data 2.1. Making data findable, including provisions for metadata	The code will be licenced under OSI approved licenses. When no other choice is preferred (e.g. when contributing to an existing ecosystem), the default license will be Apache-2.0 or EUPL-1.2 (In each case the decision for or against Copyleft and other implications can be re-addressed.). The source code will be designed, formatted and commented according to the established standards and practices of the corresponding programming languages. Additional documentation will be provided in text-based form along with the code. The identifiers, versions, licenses and contributors will be collected according to existing open source standards. The code repositories will be registered via DataCite or Zenodo.
2.2 Making data openly accessible	The source code will be made publicly available. The code will be licenced under OSI approved licenses. When no other choice is preferred (e.g. when contributing to an existing ecosystem), the default license will be Apache-2.0 or EUPL-1.2 (In each case the decision for or against Copyleft and other implications can be re-addressed.).
2.3. Making data interoperable	By using the established Github platform and adhering to each programming language's individual software design standard, interoperability with third-party development will be ensured.
2.4. Increase data re-use (through clarifying licences)	Code specifically developed within DESIR will be published on GitHub using the responsible partner's or the DARIAH ERIC's organisational account, using appropriate OSI approved licenses, see 2.2. Where existing solutions are extended, applicable platforms will be chosen on a per-case basis, and efforts will be undertaken to merge

	improvements back upstream. This will allow re-use and discoverability and constitutes the publication.
3. Allocation of resources	Publication of source code on GitHub is part of the development process and free of charge. Archiving of code repositories will be done by consortium members.
4. Data security	Source code will be archived in institutional repositories at the end of the project, on Zenodo and/or the DARIAH repository.
5. Ethical aspects	n/a
6. Other	n/a

Dataset 2: Survey conducted in WP6	
Data Summary	This survey aims to identify cross-disciplinary DARIAH communities and new core groups, considering gender and diversity as main variables. It also aims to explore to what extent is DARIAH is reaching such research communities in terms of use and access. It is expected to analyse to what extent these communities perceive DARIAH as a reliable, trustworthy and sustained infrastructure. Generated data will be the source to the empirical study of DARIAH's usage in new communities, and define new strategies and the data will be exported in xsl format. Original data will be collected from researchers defined within target groups (three cross-disciplinary communities and core groups: i) early career researchers, including MA and PhD students; ii) academics without permanent institutional affiliations (no reliable access to RIs); iii) academics with cross-disciplinary backgrounds and research interests (not clearly associated merely with one academic discipline). Data size cannot be specified yet and no specific data re-use is expected within this data collection. The collected data will be mostly useful for WP6 study. The analysed data will be, nonetheless, useful for future comparative studies.
2. FAIR Data 2.1. Making data findable, including provisions for metadata	Files will have common identifiers (respondents' identification number and country codes) and identification numbers will be consistent across all data files. Users will be given access to contextual, multilevel and thematic data.
2.2 Making data openly accessible	In accordance with data protection, only anonymous data will be available to users. No specific methods or software tools are needed to access the data to be made available.
2.3. Making data interoperable	n/a
2.4. Increase data re-use (through clarifying licences)	Analysed data and final report will be published under Creative Commons Attribution 4.0 License, allowing data re-use by third parties after the end of the project

	Data quality will be assured through quality assessment, including the quality and comparability of measurement instruments, the assessment of target-groups sample composition and the output quality of the survey.
3. Allocation of resources	The data collection and analysis is part of WP6 tasks within the project and no additional costs are expected.
4. Data security	Data will be archived in institutional repositories at the end of the project and/or the DARIAH repository.
5. Ethical aspects	Research data will be generated from the participation of humans through a survey. Respondent, which will be anonymous and will participate on a volunteer basis, will be fully aware of the nature and purpose of the research, what their role in it will be, and how the data they provide will be subsequently used. The template of the survey will has been included in D6.1 and will be provided on request.
6. Other	n/a