



Project Title	Fostering Fair Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	01 March 2019
Duration of Project	36 months
Project Website	<a href="http://www.fairsfair.eu">www.fairsfair.eu</a>

## D1.2 Data Management Plan

Work Package	WP1, Project Management and Sustainability
Lead Author (Org)	Marjan Grootveld (DANS)
Contributing Author(s) (Org)	Joy Davidson (DC/WP3), Eliane Fankhauser (DANS/PCO, WP1), Mustapha Mokrane (DANS/PCO, WP1), Sarah Jones (DCC/WP3), Brian Matthews (STFC/WP6), Jessica Parland-von Essen (CSC/WP2), Sara Pittonet (Trust-IT/WP5), Ilona von Stein (DANS/WP4), Lennart Stoy (EUA/WP7)
Due Date	31.08.2019, M6
Date	22.08.2019 <b>Draft not yet approved by the European Commission</b>
Version	1.0

### Dissemination Level

- |                                     |  |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | PU: Public   |
| <input type="checkbox"/>            | PP: Restricted to other programme participants (including the Commission)        |
| <input type="checkbox"/>            | RE: Restricted to a group specified by the consortium (including the Commission) |
| <input type="checkbox"/>            | CO: Confidential, only for members of the consortium (including the Commission)  |

### Versioning and contribution history

Version	Date	Authors	Notes
0.1	09.04.2019	Marjan Grootveld (DANS)	Generic draft
0.2	02.05.2019	Eliane Fankhauser (DANS)	PCO information added, some WP4 input added
0.3	14.05.2019	Marjan Grootveld (DANS)	FsF template; required text marked
0.4	15.07.2019	Marjan Grootveld (DANS) and all WP leaders	Edited input from the WP leaders
0.5	22.08.2019	Eliane Fankhauser and Marjan Grootveld (DANS)	Reviewed by Rob Hooft (DTL) and Heidi Laine (CSC); feedback processed

### Disclaimer

FAIRSF AIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

## Executive summary

This deliverable is the first version of the FAIRsFAIR Data Management Plan (DMP). It contains an initial description of the datasets (to be) collected, processed or generated by the project and an initial plan on how sharing, archiving and preservation of these datasets will be guaranteed.

The DMP is intended to be a living document and thus the content of this deliverable will be updated and described in more detail as the implementation of the project progresses, in order to reflect adaptations and changes with respect to the first version in month 6.

## Contents

<b>1. Introduction</b>	<b>5</b>
<b>2. Data Summary</b>	<b>5</b>
<b>3. FAIR data</b>	<b>8</b>
3.1. Making data findable	8
3.2. Making data accessible	8
3.3. Making data interoperable	10
3.4. Making data reusable	10
<b>4. Allocation of resources</b>	<b>11</b>
<b>5. Data security</b>	<b>11</b>
<b>6. Ethical aspects</b>	<b>13</b>
<b>7. Other issues</b>	<b>14</b>

## 1. Introduction

FAIRSFAR is a Coordination and Support Action (CSA) to foster FAIR data practices in Europe. FAIRSFAR will collect and generate data about the current policies, practices, services and expertise with regard to research data stewardship, with a focus on the Findability, Accessibility, Interoperability, and Reusability of these data. FAIRSFAR will also provide recommendations, including the support for organisations to take up these recommendations in their policies, practice, and/or FAIR-supporting services. Furthermore, FAIRSFAR will provide and promote FAIR competencies by means of workshops, training material for FAIR certification, data schools, a Competence Centre for research communities and FAIR competence training for higher education.

Despite the focus on FAIR research data, the project itself, being a CSA, does not intend to generate large amounts of *research* data, as that is not part of its methodology. Examples of research data that FAIRSFAR may generate include results of surveys, ontology design and data underlying the FAIR competence adoption handbook. This Data Management Plan describes how the project will handle such data.

At the same time, the project will collect a good amount of *managerial* data which is needed for running the project or for communication with external partners like research communities and research funders. Contact databases are an example. Managerial data is not included in the data described in this DMP, apart from information about personal data in Section “Ethical aspects”.

## 2. Data Summary

- What is the purpose of the data collection/generation?
- What is the relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might the data be useful ('data utility')?

In order to **foster FAIR data practices in Europe** the project will collect and analyse rich information on the current state of affairs, and deliver recommendations as well as practical approaches to improve or increase data policies, practices, services and competence. Therefore the project will collect and manage different kinds of data:

- datasets collected through, for instance, desk research, survey questionnaires, interviews, workshops and pilot studies, which will be used for analyses, recommendations, and reports.

These datasets are collections of standard material produced by a research project, e.g. project deliverables, dissemination material, training materials.

- software code to develop tools for 1) “FAIRifying” pilot repositories (Task 2.3), 2) Automated FAIR assessments of datasets (Task 4.5) and 3) a registry for FAIR repositories (Task 4.4). The latter task builds on and expands an existing service, and hence reuses existing code.

The data collected in FAIRsFAIR will be **useful** for research performing and funding organisations, for policy makers in higher education, for data service providers like digital repositories, for e-infrastructures and research infrastructures. In particular FAIRsFAIR will cooperate with the INFRAEOSC 5B projects as we design our survey. This way we can ensure that landscape analysis work in the various projects can be aligned; also the results can be shared and integrated. An agreement is currently being signed between FAIRsFAIR and the INFRAEOSC 5B projects outlining the nature of the collaboration.

More specifically, the following activities will lead to data collection, analysis, storage and publication:

1. Several tasks in WP2 (FAIR Practices: Semantics, Interoperability and Services), WP3 (FAIR Data Policy and Practice), WP4 (FAIR Certification), WP6 (FAIR Competence Centre), and WP7 (FAIR Data Science and Professionalisation) collaboratively prepare the surveys that are planned in these tasks.

The **aim of the cross-WP survey** is to get information from a maximum number of respondents while minimising the burden on them. The data collection will concern institutions rather than individuals, e.g. concerning FAIR practices (WP3), requirements for interoperability (WP2) or availability of career paths (WP7).

Furthermore, WP7 will collect data in the form of an institutional survey aiming to gather institutional information on the following aspects:

- The integration of FAIR competences in existing academic curricula in all three cycles (Bachelor, Master, PhD),
- the use of competence frameworks by the surveyed institutions
- training needs from HEIs and their staff to integrate FAIR data competences within their activities (e.g. doctoral education) and academic curricula, to support the activities under T7.4 and T7.5.

When possible, we **reuse existing information** in the preparation and extract information from public reports and documentation. For the landscape assessment WP2 and WP3 will perform desktop research about FAIR technologies, policies and practices. This will use openly accessible sources and reuse existing, openly licensed survey data wherever feasible (e.g., State of Open Data 2018, OpenAIRE survey about Horizon2020 template for Data Management Plans). The desk research results will, combined with the survey data, be published as part of the deliverables.

2. **Interview data** will be gathered in WP2. The number of interviewees and the handling of the data will be determined at a later stage, and the decisions will be based on the information needs that we have after the desktop research and survey data analysis. This approach minimises the need to ask for personal data in the survey and maximises efficiency, because interviews are time consuming for all parties involved.
  
3. **Software** will be developed to deliver and validate tools for “FAIRifying” pilot repositories (Task 2.3) and automated dataset evaluations in regard to FAIR (task 4.5), as well as a registry for FAIR repositories (Task 4.4). Task 4.4 builds on and expands an existing DataCite service, and hence **reuses existing code** which is available on Github. New code will be documented and published on Github where it can be versioned for reuse. There is a standard upload connection from Github to Zenodo, where a persistent identifier is assigned.

The **estimated volume of the data that will be archived and published** is less than 50 GB. This consists of:

Origin	File format	Methods or software needed (if any)	Estimated volume	Task number or WP number	# in the description above
Surveys	.csv	standard office software	<10 GB	WP2, 3, 4, 6, 7  Task 7.1	1
Survey documentation (e.g. questionnaire, codebook)	.txt, .pdf/A	standard office software	<10 GB	WP2, 3, 4, 6, 7  Task 7.1	1
Interviews	.txt, .pdf/A	standard office software	<10 GB	WP2, 3, 6, 7	1
Software code	t.d.b.	t.b.d.	<10 GB	T4.4, T4.5, T.2.3	2

*Table 1: list of all datasets to be generated and published by FAIRsFAIR, with volume indication*

The file formats chosen are open and comply with good practices and sustainability preferences of the repositories where we will publish the data and the software code. More information about code formats will be given in the next version of the DMP.

### 3. FAIR data

#### 3.1. Making data findable

- Are the data produced and/or used in the project discoverable with metadata?
- Are the data identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions do you follow?
- Will search keywords be provided that optimize possibilities for re-use?
- What is the approach for clear versioning?
- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

At the end of the project the data will be published in a certified digital repository, probably the EASY archive of project leader DANS (<https://easy.dans.knaw.nl/>). This will provide Dublin Core metadata, which contain keyword metadata (called “Subject”). In the repository each dataset will have a unique and persistent identifier. Furthermore, we will set up a FAIRsFAIR community space in the Zenodo archive (<https://zenodo.org/>) for project reports and software code with associated documentation (through GitHub). Zenodo, too, assigns persistent identifiers and is compliant with the DataCite metadata schema.

File names will contain the name and number of the deliverable or milestone for which the data have been (primarily) collected, plus indicators like “raw”, “version\_date” and “final”. Project deliverables based on and referring to files, as well as other documentation (see next section) will provide more context information.

#### 3.2. Making data accessible

- Which data produced and/or used in the project will be made openly available as the default?
- If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- If there are restrictions on use, how will access be provided?



- Is there a need for a data access committee?
- Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?
- How will the data be made available (e.g. by deposition in a repository)?
- 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)?
- Where will the data and associated metadata, documentation and code be deposited?
- Have you explored appropriate arrangements with the identified repository?

Anonymised and anonymous survey data with documentation such as questionnaires and codebooks, as well as documented software code will be published in certified digital repositories<sup>1</sup>. We may not be able to share interview data if an interviewee doesn't consent to sharing or withdraws their consent. In such a case these data will not be made available. Otherwise, interview data will be similarly published in a certified digital repository.

The FAIRsFAIR Consortium Agreement states in §8 that “all Results of the Project itself will be shared and placed under appropriate licenses that allow re-use with attribution and that prohibit the subsequent application of more restrictive licensing.” We will therefore apply a CC BY-SA licence to all data that can be published (see previous section about interview data) and a GNU General Public License 3 (GPL-3.0) type of licence to software<sup>2</sup>. The project results will be made available in a certified digital repository that supports this type of licence (see also section “Making data findable”). As can be gathered from the table in section “Data summary”, probably no specific methods or software tools are needed to access the data; for code this will be established later.

The EASY archive of project leader DANS (<https://easy.dans.knaw.nl/>) will be used to preserve the data. This is a CoreTrustSeal-certified repository for research data, which accommodates both Open Access with the CC licence, and Restricted Access for the interview data with a stricter licence, for only those users who receive permission to access these data. All metadata in EASY are public. They adhere to the Dublin Core metadata standard<sup>3</sup> and implementation controlled vocabularies are

---

<sup>1</sup> The template text for our project surveys states among other things: “All the collected data will be anonymised (...) The raw data will be stored on the internal servers of FAIRsFAIR project partners, protected by passwords only known to researchers conducting the survey. (...) The information you provide will be analysed and presented in aggregate form in project reports together with the information from other participants (...) Anonymised and aggregated data will be stored in a certified data repository after the end of the project and be made available for use by third parties.”

<sup>2</sup> In addition, all planned project deliverables from Work Packages 1-7, including this DMP, will be public.

<sup>3</sup> Dublin Core Metadata Initiative (DCMI) specifications: <https://www.dublincore.org/specifications/dublin-core/>

used for e.g. media type, language, and the relation of the dataset to other resources. EASY is a long-term archive and preservation and availability are in principle “indefinitely”.

### 3.3. Making data interoperable

- Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
- What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
- Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mappings to more commonly used ontologies?

The file formats we will deliver (see table above) are open and allow straightforward reuse. The Dublin Core and DataCite metadata specifications are widely accepted and implemented.

Furthermore, interoperability is the express ambition of WP2 (“FAIR practices: Semantics, Interoperability, and Services”), which will actively participate in domain-specific and cross-disciplinary initiatives involved in semantic interoperability.

### 3.4. Making data reusable

- Are data quality assurance processes described?
- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If applicable, specify why and how long a data embargo is needed.
- Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?

To the extent that data form the basis of project deliverables, internal quality review procedures will apply. These are described in the Project Handbook (D1.1). Other quality aspects, such as an explicit methodology, are covered by the FAIR aspects addressed above, and in particular by providing sufficient documentation to interpret the data and the code correctly, which relates to FAIR

principle R1.2: “(Meta)data are associated with detailed provenance”. Such documentation, e.g. sample questionnaires, codebooks and templates of informed consent forms (see Section “Ethical aspects”), will be published along with the data and code.

The FAIRsFAIR Consortium Agreement states in §8 that “all Results of the Project itself will be shared and placed under appropriate licenses that allow re-use with attribution and that prohibit the subsequent application of more restrictive licensing.” We will therefore apply a CC BY-SA licence to data and a GNU General Public License 3 (GPL-3.0) type of licence to software.

Data and code will be made available through certified digital repositories which supports the types of licences and which will preserve the data for the long term, in principle “indefinitely”. We will set up a FAIRsFAIR community space in the Zenodo archive (<https://zenodo.org/>) for the project results; at the end of the project this community space could be handed over to an EOSC organisation for future extension and maintenance, ownership arrangements permitting.

## 4. Allocation of resources

- What are the costs for making data FAIR in your project?
- How will these be covered?
- Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
- Who will be responsible for data management in your project?

Delivering data that are as FAIR as possible is on the project team’s mind. Therefore we don’t plan a separate step or budget for making data FAIR. The amount of research data envisaged in the project is very modest, therefore a cost/benefit analysis for the long-term storage of each component is not needed. The costs for long-term - in principle “indefinite” - preservation in a trustworthy archive are covered.

Each work package leader is responsible for data management in their WP, including the implementation of and, if necessary, updates of this DMP. The Project Coordination Office is overall responsible for data management and for evaluating the implementation of this DMP. Evaluations will take place in months 20 and 34.

## 5. Data security

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long term preservation and curation?

**Near the end of the project** the research data and code will be published and safely preserved in a certified digital repository, with good and transparent data security processes in place. The CoreTrustSeal certification requirements include a pertaining requirement: “The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users”<sup>4</sup>.

**During the project** the research data are stored in a FAIRsFAIR Google Team Drive, with folders per Work Package; see D1.1 Project Handbook, Chapter 4.2. This allows for file sharing across partners and keeping track of revisions. The project uses the Google Vault. A retention rule is set for the project to keep indefinitely files stored in Team Drive and Gmail (including after deleting and emptying the trash).

Access to the Google Team Drive is managed by the Project Coordination Office. Google’s network is protected from external attacks. Data belonging to G Suite customers is stored at rest in two types of systems: disks and backup media. Google also stores data on offline backup media to help ensure recovery from any catastrophic error or natural disaster at one of their data centers. G suite offers a data loss prevention policy to protect sensitive information within Gmail and Drive. With respect to data protection, Google is committed to complying with the EU General Data Protection Regulation (GDPR) for G Suite<sup>5</sup>.

For conducting and analysing surveys FAIRsFAIR will use EUSurvey<sup>6</sup>, Qualtrics<sup>7</sup> or Online Surveys<sup>8</sup>. Qualtrics and Online Surveys have the ISO 27001 certification for an Information Security Management System<sup>9</sup>. All three survey tools are GDPR-compliant.

During the project Work Packages will also use secure institutional services for storing data. For example, Work Package 3 will make use of DataStore<sup>10</sup> at the University of Edinburgh to securely store the survey data. DataStore is fully backed-up, secure and resilient. We anticipate collecting a modest amount of data and foresee no problems with using DataStore as it can accommodate data of more than a petabyte. DataStore allows selected collaborators to access the data via DataSync which is a Dropbox-like file sharing service for academic institutions. The Datasync service is hosted entirely within the University of Edinburgh and no data is stored on third-party servers. By default,

---

<sup>4</sup> [https://www.coretrustseal.org/wp-content/uploads/2017/01/Core\\_Trustworthy\\_Data\\_Repositories\\_Requirements\\_01\\_00.pdf](https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf) R16

<sup>5</sup> [https://cloud.google.com/security/infrastructure/design/resources/google\\_infrastructure\\_whitepaper\\_fa.pdf](https://cloud.google.com/security/infrastructure/design/resources/google_infrastructure_whitepaper_fa.pdf) and <https://cloud.google.com/security/compliance/gdpr/>

<sup>6</sup> EUSurvey: <https://ec.europa.eu/eusurvey/> and <https://ec.europa.eu/eusurvey/home/helpauthors>

<sup>7</sup> Qualtrics: <https://www.qualtrics.com/> and <https://www.qualtrics.com/uk/platform/gdpr/> and <https://www.qualtrics.com/privacy-statement/> and <https://www.qualtrics.com/platform/security/>

<sup>8</sup> Online Surveys: <https://www.onlinesurveys.ac.uk/> and <https://www.onlinesurveys.ac.uk/gdpr/> and <https://www.onlinesurveys.ac.uk/help-support/online-surveys-security/>

<sup>9</sup> <https://www.iso.org/isoiec-27001-information-security.html>

<sup>10</sup> <https://www.ed.ac.uk/information-services/research-support/research-data-service/during/data-storage>

none of the data will be shared by DataSync<sup>11</sup> until we deliberately choose to share it with specific individuals. During the active stage of the data collection and analysis Work Package leader DCC will control the sharing of the raw survey data and interview data with external partners as agreed with the FAIRsFAIR Executive Board.

In Work Package 5 - “Engagement, communication and uptake” - WP leader Trust-IT stores data on the organisation’s networks via the following providers: Amazon AWS, It.NET and Hetzner server farm, which ensures the high standard security access police. All the virtual machines are protected by firewalls where the website and the database reside. Daily backups, snapshots, periodical security updates on the website and on the machine operating system are ensured. Trust-IT also uses a series of monitoring tools to prevent DOS attacks and to verify the integrity of the machines.

## 6. Ethical aspects

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
- Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

FAIRsFAIR reaches out to many organisations, individuals, and other projects, and will organise surveys, interviews, workshops and other meetings, and repository calls. Data collected for the purposes of administering the project activities will be held securely and according to legislation, and will not be shared externally. We will only collect personal data that is necessary to fulfill the information needs of the project (respecting the principle of data minimisation), and we will not collect “special categories of data” in terms of the GDPR<sup>12</sup>. Although administrative personal data are outside the scope of the DMP, we describe here how they are managed:

Any personal data will be protected, as stated in Article 39 of the Grant Agreement. FAIRsFAIR has a [Privacy Policy Statement](#), which addresses personal data (processing, data subject’s rights, opt-out, cookies used on the website and in social media, et cetera). For example, personal data processed for applications collected via the FAIRsFAIR website will be kept by DANS-KNAW as Data Controller in terms of the GDPR for up to 5 years, to allow for possible external audits, as requested by contractual provisions the Data Controller is subjected to. In addition to the Privacy Policy Statement there are [Terms of Use](#) for services provided via the project website.

The project leader DANS-KNAW has appointed a Data Protection Officer. The contact details of the Data Protection Officer are made available to all data subjects involved in the research.

Survey, interview and workshop participants (Work Packages 2, 3, 4, 6, 7) as well as repositories responding to the certification support call (Tasks 2.3 and 4.2) will be provided with a clear

<sup>11</sup> <https://www.ed.ac.uk/information-services/research-support/research-data-service/during/collaboration>

<sup>12</sup> <https://gdpr-info.eu/art-9-gdpr/>

statement on the purposes of the data collection, how the data will be used, and with whom it may be shared<sup>13</sup>. Participants will have the opportunity to decide if they want to provide any personal information such as their name and email address. Those who choose to provide us with this information and agree to be contacted for interviews or for news and updates on the project work will be added to the contacts database being managed by Trust-IT, as Data Processor in terms of the GDPR.

We developed a project-wide template for informed consent with respect to interviews. The template can only be used in combination with an information sheet (in language and terms intelligible to the participants). Detailed information on the informed consent procedures in regard to data processing will be kept on file and be archived. Also, templates of the informed consent forms and information sheets will be kept on file and be archived.

Project deliverable 8.1, “POPD - Requirement No. 1”, which is due in month 36, will describe the ethics requirements that the project deals with.

## 7. Other issues

- Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

FAIRSFAR partners agree with the principles and good practices in the European Code of Conduct for Research Integrity (ALLEA, 2017<sup>14</sup>). In addition, project partners may have their own national and/or institutional data management policies for data in their charge.

---

<sup>13</sup> From the template privacy statement for our project surveys: “Data on opinions, attitudes and practices of the respondents related to the survey themes will be gathered. Personal data (including name/surname, an e-mail address and the name of the respondent’s organisation) is needed to implement the survey. This data will be used only to contact you only if you choose to allow this. All data is stored on the EU server site, which is password protected and only staff working with the survey can access it. All the data gathering activities conducted in the framework of this survey will comply with the requirements of GDPR. (...) FAIRSFAR does not transmit any personal data to other parties. All the collected data will be anonymised, removing indirect identifiers such as your role and name of organisation, and treated confidentially. The information you provide will be analysed and presented in aggregate form in project reports together with the information from other participants, and hence no personal information will be revealed. The raw data will be stored on the internal servers of FAIRSFAR project partners, protected by passwords only known to researchers conducting the survey. All the raw data will be deleted after the project is finalised. The anonymised and aggregated data will be stored in a certified data repository after the end of the project and be made available for use by third parties.”

<sup>14</sup>

<https://www.allea.org/wp-content/uploads/2017/04/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf>