Deliverable D1.1b

Data Management Plan

Making Software FAIR: A machine-assisted workflow for the research software lifecycle

Project acronym: SoFAIR

Grant Agreement Number: CHIST-ERA-22-ORD-08

Deliverable information	
Deliverable number and name	D1.1b Data Management Plan
Due date	M3
Actual delivery date	
Work Package	WP1
Lead Partner for deliverable	The Open University
Authors	David Pride, Petr Knoth
Reviewers	Pavel Smrž, Laurent Romary
Approved by	
Dissemination level	Public
Version	0.1

Document revision history			
Issue Date	Version	Author	Comments
19/02/2024	0.1	David Pride, Petr Knoth	First draft
25/02/2024§	Final	David Pride, Petr Knoth	Final version for submission

Table of Contents

Tables

Abbreviations

DMP: Data Management Plan

Executive summary

This Data Management Plan (DMP) contains a description of all datasets to be collected, generated or re-used as part of the SoFAIR project's research activities. This includes the purpose of data collection, data origin, type, sharing and utility.

This DMP outlines measures taken to ensure compliance with the FAIR data principles. Most datasets generated during SoFAIR will be made findable and openly accessible (whenever possible, depending on data protection requirements). To make data interoperable, the consortium will use Open Source software and file formats wherever possible. To enable and stimulate re-use of data, analysis code will be shared via GitHub, appropriate open licences will be assigned, and quality is ensured via internal review processes. We plan to use CC-BY unless other onctraints force us not to do so.

The DMP further describes the allocation of resources for data management, storage solutions and back-up plans during the project, long-term retention and data security, ensured via encryption and password protection. It finishes with a description of technical and organisational measures put in place for the protection of personal data, including collection and internal sharing of personal data (where necessary).

1. Introduction

This data management plan (DMP) articulates the SoFAIR consortium's strategy for the stewardship of scientific outputs generated throughout the project. It outlines measures taken to ensure compliance with the FAIR data principles and describes the allocation of resources for data management, storage solutions and back-up plans during the project. By implementing FAIR compliant practices and open access policies, the plan aims to ensure that the project's outputs are findable, accessible, interoperable, and reusable. The commitment to sharing data and results in a structured digital format is intended to support the advancement of research and foster a collaborative scientific environment. This approach aligns with the consortium's objectives to contribute to the scientific community and enhance the utility of its research outputs.

2. Data summary

SoFAIR will follow the overarching principles listed below:

- All publications will be made available under specific OA licences with minimal restrictions.
- Data will be FAIR by design and made available according to the principles “as open as possible” and “as early as possible”
- Services and tools will be based on open source software/solutions
- All documentation and technical specifications will be available in relevant open repositories, such as Zenodo.

The specific outputs of the project, as described in the DoW, will consist of the following:

- *Publications*: Project reports, summaries and academic outputs will be produced in PDF format and stored in FAIR compliant repositories offering PIDs, standard metadata and robust accessibility. Documents tagged as “public” will be shared in open access with permissive Creative Commons licences,
- *Data:* the project will produce 1) gold standard annotated datasets needed for further production of tools and services aiming to automatically extract software mentions and disambiguate them. This will be valuable corpora for future research and practice; 2) language models for software mentions recognition; the annotated corpora and language models will be made available as openly as possible and in FAIR compliant repositories;
- *Software:* Where possible, software developed will be open source or part of open source software, extensively documented and released under clear and liberal open source licences to enable maximum reuse by the communities. These outputs will be available on widely accessible repositories (e.g. Github), which in turn are harvested and preserved within the Software Heritage infrastructure.
- *Training materials*: All training materials will be properly described with appropriate metadata and published as Open Access.

We are committed to making all scientific outputs, including data, software, and publications, openly and freely accessible within open repositories.

Partners will voluntarily share their research data, which include results, models and datasets. These data will be shared in a digital and structured form, to facilitate and encourage other researchers to freely access, mine, exploit, reproduce and disseminate our data and results.

3. FAIR data

The following sections describe which steps will be taken to make the data findable, (openly, where possible) accessible, interoperable and re-usable.

3.1. Making data findable, including provisions for metadata

Raw data and all processed versions of data will be saved in separate folders. Dataset names will include version numbers, starting with v01 for the raw data. Datasets will be named according to the following convention, using the project name, the work package number, title, version number, year and month. All publications will be stored in FAIR compliant repositories and will be assigned identifiers.

3.2. Making data openly accessible

Whenever possible, open datasets will be shared in a format accessible with open source software. No proprietary software is needed to re-use our data, and all used proprietary software is commonly used for research purposes.

3.3. Making data interoperable

To allow data exchange and re-use, all data will be saved in the most interoperable format possible, such as .txt or .pdf for text data, .csv for tabular data, .R and .py for code, .tif, .png, .svg, .jpeg for images.

3.4. Increase data re-use

Data will be made available for re-use with a licence not more restrictive than CC-BY. By default, unless there is a rational reason otherwise, the data produced by the project will be fully available for others to re-use.

3.4.1. Procedures for quality assurance

The SoFAIR consortium will ensure that all shared data is consistent with quality standards required to publish in peer-reviewed journals. To the best of our abilities, we will describe the data collection and analysis procedures in sufficient detail to allow reproducibility by trained professionals in the respective project reports or research papers. All software code developed for this project will be hosted on https://github.com/SoFairOA

As a general measure of quality assurance, all analyses and deliverables undergo an internal review process involving at least two assigned consortium members.

4. Allocation of resources

The primary responsibility for data management for SoFAIR is with the project coordinator. However, each task leader is responsible for handling the data they collect as part of their tasks according to the regulations and following the procedures laid out in this DMP.

5. Data security

During the project, a copy of all data acquired by consortium members will be maintained by the partner/s responsible for the respective task. All data will be stored on servers or within services that are backed up, are located within the infrastructure of the project partners, or within cloud services that are certified for use by the IT departments of the project partners.

6. Protection of Personal Data

6.1. Definition of personal data

The SoFAIR consortium refers to the GDPR definition of personal data as follows: “Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data. Personal data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person remains personal data and falls within the scope of the GDPR. Personal data that has been rendered anonymous in such a way that the individual is not or no longer identifiable is no longer considered personal data. For data to be truly anonymised, the anonymisation must be irreversible.”[1] Examples of personal data are a name and surname; a home address; an email address such as name.surname@company.com; an identification card number; location data; an Internet Protocol (IP) address; a cookie ID; the advertising identifier of a phone. In contrast, data such as a company registration number, and anonymised data are not considered personal data. Additionally, the GDPR does not apply to information “about legal entities such as corporations, foundations and institutions. […] Data must therefore be assignable to identified or identifiable living persons to be considered personal.”[2]

6.2. Collection and sharing of personal data in SoFAIR

The SoFAIR project’s data collection and management activities will strictly adhere to the General Data Protection Regulation (GDPR) guidelines. The project excludes the collection, processing, or storage of personal data. This deliberate approach ensures compliance with legal standards for data privacy.

[1] https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

[2] https://gdpr-info.eu/issues/personal-data

As part of the project, the email addresses of authors within publicly available research papers will be processed. These email addresses may then be used by the authors' respective institutions, i.e. those institutions that the authors are affiliated with, for the purposes of checking whether new software identified in the author's manuscript should be registered and archived for future generations. These email addresses were made available within the manuscripts by the authors specifically for the purposes of being contacted with respect to this research, which is exactly the only kind of contact the project will facilitate. The processing also falls under the GDPR exception for Research and Statistics as the SoFAIR project's research uses rigorous scientific methods and furthers a general public interest. This right exempts the project from several of the GDPR's provisions on: the right of access; the right to rectification; the right to restrict processing; and the right to object.

7. Other institutional regulations

All partner institutions require their employees to enter publications to an internal publication database or repository. In addition, the partners listed below have to follow other institutional procedures, which are in accordance with the present Data Management Plan.