Grant Agreement Number: 957204 (H2020-ICT-38-2020)
Project Acronym: MAS4AI
Project Start Date: 1st October 2020
Project Full Title: Multi-Agent Systems for Pervasive Artificial Intelligence for assisting Humans in Modular Production

D8.2 – Data Management Plan

Dissemination level:	PU
Date:	2021-03-17
Deliverable leader:	DFKI
Contributors:	DFKI
Reviewers:	US, TNO
Type:	ORD P
WP / Task responsible:	8
Keywords:	FAIR principles, data management, GDRP

Executive Summary

The Data management plan aims give an overview on how data will be generated, collected, documented and saved within in the MAS4AI project.

Document History			
Version	**Date**	**Contributors**	**Description**
V0.1	2021-03-17	Jens Popper	Initial document
V1.0	2021-03-31	Jens Popper	V 1.0
V2.0	2022-03-09	Aleksandr Sidorenko	V 1.0 Update

Table of Contents

Table of Figures

1 Introduction

The Data management plan aims give an overview on how data will be generated, collected, documented and saved within in the MAS4AI project. It relies on the Guidelines of FAIR Data management in Horizon 2020, which describes it as a key element that ensures how data is properly managed. A basic differentiation of the data collected in MAS4AI can be made between

- Data generated from open accessible information, such as reports on multiagent systems, testbed descriptions, statistics of relevant industry sectors and available setup systems that might be relevant for the MAS4AI project,
- Data generated by partners of the project and evaluator activities, such as deliverables, meeting minutes, technical documents, guidelines, whitepapers and other work created in the MAS4AI project.

Types of data, that can be collected to achieve the documents that are being described in the text above, include:

- Personal data
- Technical data
- Business-related data

All types of data collected should serve to advance the objectives of the project, such as online collaboration within the project consortium or with external partners such as the advisory board, for validation and testing activities, or to advance the evaluation and exploitation of the project results.

2 Versions and Updates

The data management plan will be updated during the duration of the project, as stated by the Guidelines of FAIR data management of Horizon 2020. These updates might occur due to:

- Significant changes in the project, such as the availability of essential new data sources or changes in the test bed for the MAS4AI system
- Changes in consortium policies
- Changes in the project consortium
- Other reasons that might be of relevance for the project

To keep up to date with changes, the data management plan will be reviewed and revised as is stated in Table 1.

Table 1 Data management plan timetable

Version	Content	Delivery due
V1	The first version of the data management plan will take into account the requirements of the General Data Protection Regulation (GDPR). The GDRP was put into affect on May 25, 2018 and lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the movement of personal data.	M6
V2	Version 2 will update the data management plan at the mid-term of the project	M18
V3	Version 3 will present a final version of the data management plan. As such, it	M36

	will be done before the end of the project and contain information on how data will be administered once the project has finished.	

3 Open Dissemination Strategy

The MAS4AI consortium will follow an open dissemination strategy to make the project results available for the public. Project partners will be notified of any dissemination activities with sufficient information of the results to be publicized. Research partners in the project will aim to publish in open access publications and are provided with dedicated budgets to publish according to the “Gold Model” standard (publication with immediate, free of charge open access). In cases, in which the “Gold Model” is not applicable (such if the dissemination budget reserved for open access is exhausted), the open access will be given according to the “Green Model” (publications are made available on a repository on the project website) after a possible embargo period.

4 Protection of Personal data

Collection of personal data will be reduced to minimum, and only if it is necessary to advance the project objectives. Such personal data might be necessary to establish the required collaboration among the project partners and external partners, such as the advisory board. All personal data that is collected in the MAS4AI project will be securely stored.

All personal data protection processes will follow the Regulation (EU) 2016/676 of the European Parliament and of the of the Council on the protection of natural persons regarding the processing of personal data and on the free movement of such data as well as Directive 95/46/EC of the GDRP.

Personal identifiers will be protected through pseudonymization. Following Article 4 paragraph 1 of the GDRP, personal data contains all information that might be used to identify a person directly or indirectly, such as identifiers as name, identification number, location data, online identifiers or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Relations between persons and codes will be known only to the project partners that are responsible for the secure storage of the records.

In situations, in which it will be necessary to transmit data to entities outside of the EU, the MAS4AI project will apply the transfer tools provided for in Art. 44 - 49 GDPR. Especially the transfer may take place to countries that have been recognized by the European Commission as providing adequate protection in terms of the protection of privacy and fundamental rights and freedoms of individuals.

5 FAIR Data Strategy

The "FAIR Guiding Principles for scientific data management and stewardship", as defined in [1], intend to provide guidelines to improve the Findability, Accessibility, Interoperability and Reuse of digital assets. It specifies its guidelines as follows:

- Findable
  - F1 (Meta)data are assigned a globally and persistent identifier
  - F2 Data are described with rich metadata
  - F3 Metadata clearly and explicitly include the identifier of the data they describe
  - F4 (Meta)data are registered or indexed in a searchable resource
- Accessible
  - A1 (Meta)data are retrievable by their identifier using a standardised communications protocol
    - A1.1 The protocol is open, free, and universally implementable
    - A1.2 The protocol allows for an authentication and authorisation 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 FARI principles
  - I3 (Meta)data include qualified references to other (meta)data
- Reusable
  - R1 (Meta)data are richly described with a plurality of accurate and relevant attributes
    - R1.1 (Meta)data are released with a clear and accessible data usage licence
    - R1.2 (Meta)data are associated with detailed provenance
    - R1.3 (Meta)data meet domain-relevant community standards

The MAS4AI project will follow these guiding principles for scientific data management and stewardship throughout the project.

5.1 How will data be made findable

5.1.1 Project Sharepoint

Data of the MAS4AI projects will be shared mainly through the projects' Sharepoint collaborative platform. It is protected by a 2-factor authentication process, as well as an actively managed whitelist. Thus, only authorized persons will be able to access the data on the platform. Any other ways of data transferring, such as files added to emails, will be reduced to the minimum necessary. Data that is directed to the public will be mainly shared on the projects' web presence, social media and other dissemination activities. All public project data will ensure that it will not contain critical personal data or otherwise confidential information.

Essential for data discovery is the identification of the data. To ensure that data can be identified, all documents will be assigned a unique and persistent identifier that will be given at the time of the submission process. It will contain the project acronym, the deliverable with corresponding number, month of delivery, title of the delivery and the current version of the document. Other documents will as well as deliverables contain the project acronym, the work package or task number, as well as a descriptive name with a version number of the current document.

For descriptive names for documents, the project will rely on a list of rules as described in [2]:

- Keep file names short, but meaningful
- Avoid unnecessary repetition and redundancy in file names and file paths.
- Use underscores to delimit words
- When including a number in a file name always give it as a two-digit number, i.e. 01-99, unless it is a year or another number with more than two digits.
- If using a date in the file name always state the date 'back to front', and use four digit years, two digit months and two digit days: YYYYMMDD or YYYYMM or YYYY or YYYY-YYYY.
- When including a personal name in a file name give the family name first followed by the initials.
- Avoid using common words such as 'draft' or 'letter' at the start of file names, unless doing so will make it easier to retrieve the record.
- Order the elements in a file name in the most appropriate way to retrieve the record.
- The file names of records relating to recurring events should include the date and a description of the event, except where the inclusion of any of either of these elements would be incompatible with rule 2.
- The file names of correspondence should include the name of the correspondent, an indication of the subject, the date of the correspondence and whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule 2.

- The version number of a record should be indicated in its file name by the inclusion of 'V' followed by the version number and, where applicable, 'Draft'.
- Avoid using non-alphanumeric characters in file names.

To keep track of document versioning, all templates created in the MAS4AI projects will contain a table to document the history of changes. It contains the current version of the document, the date of change, the name of the contributor as well as a short description of the changes. This table will look as shown in Table 2.

Table 2 Document history table

Document History			
Version	**Date**	**Contributors**	**Description**
V0.1	2021-03-17	Jens Popper	Initial document
V1.0	2021-03-17	Jens Popper	Final Version

The version of the document shall be noted by "VX.X" notation, where the "X.X" presents a numerical value between 0 and 1. The date shall be noted as YYYYMMDD. Only the principal author will change the draft number and will add the term "Final version" to documents that are ready to be sent to the EU or used internally.

The documents of the MAS4AI project will contain basic metadata to facilitate the efficient use of the document related information by the consortium. Furthermore, it will contribute to the discoverability and identification of relevant information contained in the document. This metadata will be included on the front page of the documents, containing information about the dissemination level, date, deliverable leader organisation, contributor organisations, reviewer organisations, document type, corresponding WP or Task and a short list of descriptive keywords. An example is given in Table 3.

Table 3 Document metadata

Dissemination level:	PU
Date:	2021-03-17
Deliverable leader:	DFKI
Contributors:	DFKI
Reviewers:	TNO, US
Type:	ORD
WP / Task responsible:	8
Keywords:	FAIR principles, data management, GDRP

The dissemination level of the document is listed in three possible types (Public, Restricted, Confidential). The Type of the document has the following four options, given with the corresponding abbreviation:

- Report (R)
- Software (SW)
- Demonstrator (D)
- Other (O)
- Open Research Data Plan (ORD P)

5.1.2 Zenodo Platform (WIP)

For sharing the project's results we will use Zenodo platform that fully follows the FAIR principles.

To make data findable Zenodo does the following:

- A DOI is issued to every published record on Zenodo. (F1)
- Zenodo's metadata is compliant with DataCite Metadata Schema[1] minimum and recommended terms, with a few additional enrichments (F2)
- The DOI is a top-level and a mandatory field in the metadata of each record. (F3)
- Metadata of each record is indexed and searchable directly in Zenodo's search engine immediately after publishing. (F4)
- Metadata of each record is sent to DataCite servers during DOI registration and indexed there. (F4)

To make data accessible:

- Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name. (A1)
- Metadata is also retrievable through the public **REST API** (A1)
- Metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it. (A1.2)
- Data and metadata will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. (A2)
- Metadata are stored in high-availability database servers at CERN, which are separate to the data itself. (A2)

[1] https://schema.datacite.org/

To make data accessible:

- Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML. (I1)
- For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE). (I2)
- Each referenced external piece of metadata is qualified by a resolvable URL. (I3)

To make data accessible:

- Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments. (R1)
- License is one of the mandatory terms in Zenodo's metadata, and is referring to an Open Definition license. (R1.1)
- Data downloaded by the users is subject to the license specified in the metadata by the uploader. (R1.1)
- All data and metadata uploaded is tracable to a registered Zenodo user. (R1.2)
- Metadata can optionally describe the original authors of the published work. (R1.2)
- Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available. (R1.3)

5.1.3 MAS4AI Community on Zenodo Platform

A MAS4AI community was created on the Zenodo platform. The screenshot of the community profile with the first uploaded paper is shown on Figure 1. All the project partners will use the following links to upload and curate their results on Zenodo platform. This will also be a part of the dissemination activities from the WP7.

Collection URL:

https://zenodo.org/communities/mas4ai/

Above address links directly to your community collection.

Upload URL:

https://zenodo.org/deposit/new?c=mas4ai

Above address will automatically ensure people who use it will have their record added to your community collection.

Curation URL:

https://zenodo.org/communities/mas4ai/curate/

Above address links to your private curation URL. You will find all uploads pending your curation.

Harvesting URL:

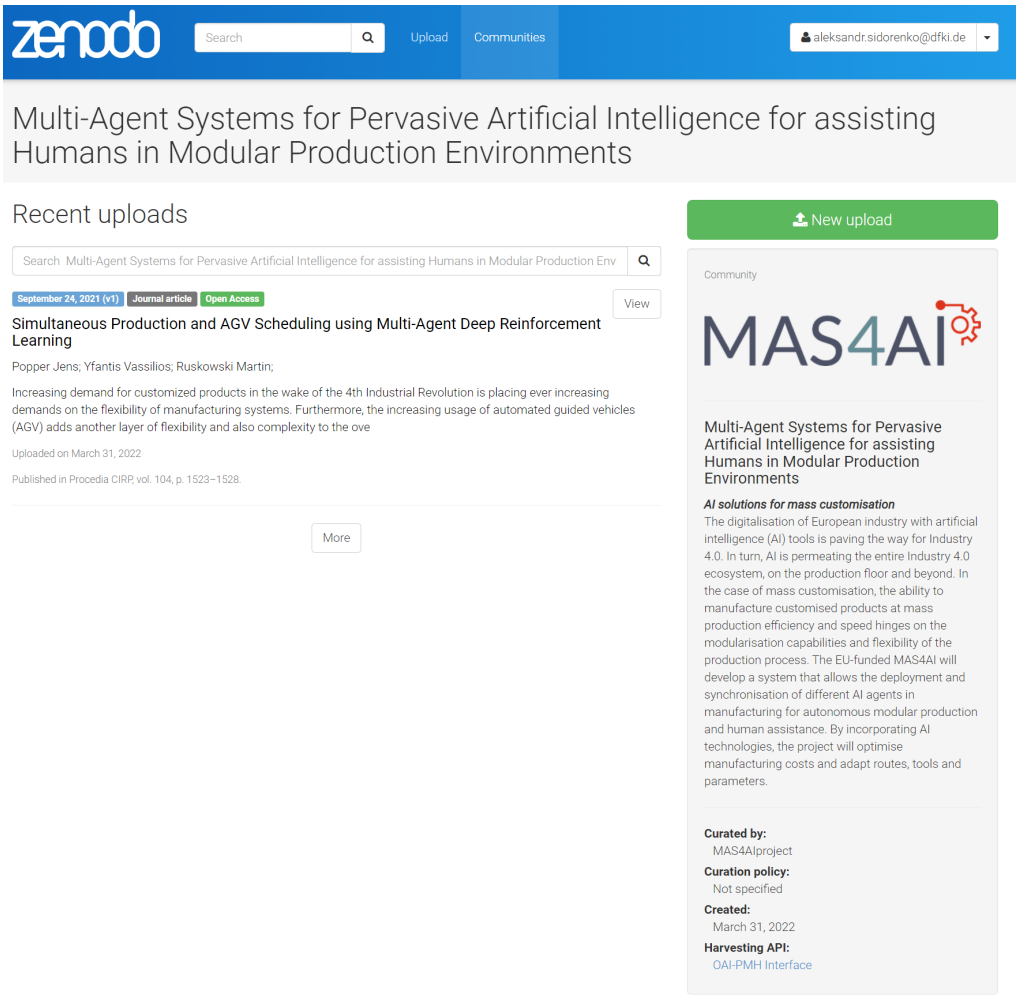

Figure 1 Zenodo MAS4AI community profile

https://zenodo.org/oai2d?verb=ListRecords&set=user-mas4ai&metadataPrefix=oai_dc

Above address links to a OAI-PMH feed, which can be used by other digital repositories to harvest this community.

5.2 How will data be made accessible?

5.2.1 Which data will be made openly available?

Project data and deliverables that are declared as public will be anonymised and made available to the public and all third parties that request access to them. Furthermore, the anonymised data sets will be exploited through the creation of tables and infographics that may be used as part of the dissemination activities of the project. Data that will be made public are technical documents required to recreate or implement the project results, such as general information on created architectures or specifications, as well as anonymised datasets for result recreation.

However, not all data will be made openly available. This applies to documents where the access needs to be restricted. The project has the necessary tools to handle data with multiple levels of restrictions associated with it. This allows for the creation of users with different access levels, access for specific time frames, access to specific data sets as well as access through secure authentication.

5.2.1.1 SmartFactoryKL Industrie4.0 Modular Testbed

5.2.1.1.1 Nature of the published data

SmartFactoryKL uses data from the abstracted manufacturing use-case developed as a laboratory scale production line. All the information and data models used in the setup are open and publicly available. Data is mostly synthesized and simulated, no real manufacturing process specific information is being used.

5.2.1.1.2 Data openly available on Zenodo Platform

All the scientific results in form of conference and journal papers with the open access status will be published on the Zenodo platform. The raw data used for training the models as well as the trained model can also be published on the platform.

5.2.1.1.3 Data available only for the MAS4AI consortium partners

Information will be openly available for usage.

5.2.1.1.4 Data not openly available

Information will be openly available for usage.

5.2.1.2 Volkswagen PKW

5.2.1.2.1 Nature of the published data

Volkswagen provides data from an abstracted manufacturing use case in a modular production system. The information used to set up the modular production system is based on openly available information. The modular production system is depicted in a simulation using Tecnomatix PlantSim with information on process times, resources and orders/products. The data is used for coordinating the processes, products and resources on relevant agents. No Volkswagen specific production data is used in the use case.

5.2.1.2.2 Data openly available on Zenodo Platform

The raw data of the use case from the simulation can be made openly available. The developed interaction models and concepts for agent collaboration and results should be accessible for consortium members only.

5.2.1.2.3 Data available only for the MAS4AI consortium partners

The developed interaction models and concepts for agent collaboration and results should be accessible for consortium members who are participating in the VW use case.

5.2.1.2.4 Data not openly available

5.2.1.3 VDL Industrial Modules

5.2.1.3.1 Nature of the published data

VDL provides data related to production planning of complex assemblies and to delivery of customized welding instructions in an adaptive way depending on the experience and skills of workers. The data needed for planning includes the individual work orders, their dependency on different shopfloor assets and the required capacity of those assets. The data linked to the adaptive operator support agent includes digital instructions of different complexity and measurements of the product quality and the actual production time per work order. Since the data is linked to proprietary information of VDL's customers and personal information of its workers it needs to be carefully curated before being publicly released.

5.2.1.3.2 Data openly available on Zenodo Platform

The information on the work orders will be anonymized by removing references to actual parts (and replacing them, in a consistent manner with e.g. "Part 1", "Part 2") to avoid infringement on the IP of the VDL's customers. Similarly, the exact names and models of the shopfloor assets of VDL will be replaced with labels like "Machine 1", "Machine 2". The links

between the work orders and the required capacities and capabilities will be maintained allowing for testing different planning and scheduling solutions.

Similarly, in the data used for the operator support the references to the actual part numbers and drawing will be replaced with new labels. Any references to the personnel IDs will also me anonymized. The released data will therefore include the work orders, measurements of production time and quality checks and the actual digital instructions (e.g. Composer files) without references to the official part numbers.

5.2.1.3.3 Data available only for the MAS4AI consortium partners

The raw data without any additional preprocessing will be made available to the MAS4AI partners, who are bound by the confidentiality clauses of the Consortium Agreement.

5.2.1.3.4 Data not openly available

All the data needed to successfully implement the MAS4AI project will be made available as described above.

5.2.1.4 Baltik Vairas

5.2.1.4.1 Nature of the published data

Baltik Vairas will publish the dataset of the painting line workload and the planning output of the agent. The workload data were generated randomly from a generation algorithm developed in MATLAB (by LMS) and designed to create artificial workload scenarios close to the actual production workload. The planning results (output of the datasets) were generated from the planning agent software that was developed in WP4 by LMS.

The workload data provide the following type of information:

- A list of production orders
- Bill of materials for each production order
- Quantity of products for each order
- Colour requirements for each order
- Size of the items with respect to the painting line carriers
- Setup delays across different colour changes

The planning output includes the following information:

- A sequence of items entering the painting line

5.2.1.4.2 Data openly available on Zenodo Platform

Baltik Vairas and LMS will publish the previously described data on ZENODO, KAGGLE or any other platform selected by the consortium

5.2.1.4.3 Data available only for the MAS4AI consortium partners

Information will be openly available for usage.

5.2.1.4.4 Data not openly available

Information will be openly available for usage.

5.2.1.5 Fersa

5.2.1.5.1 Nature of the published data

FERSA will publish their product nominal dimensions (D, d, T). FERSA is considering .CSV but the format can be discussed and changed.

5.2.1.5.2 Data openly available on Zenodo Platform

FERSA will publish the previously described data on ZENODO, KAGGLE or any other platform selected by the consortium

5.2.1.5.3 Data available only for the MAS4AI consortium partners

FERSA will give access to the consortium partners to his repository on a case by case basis. Only partners that provide inputs and/or contributions to the use case will be invited.

5.2.1.5.4 Data not openly available

FERSA will not publish any data related with their production as it is their know how on design and production can be reached which is extremely dangerous.

5.2.1.6 SCM

5.2.1.6.1 Nature of the published data

SCM will make public two datasets developed by AIMEN, one for the orchestrator agent (WP4) and another for the cutting agent (WP5). The dataset related with WP4 will contain synthetic data building a set of machine parameters for a collection of virtual machines added to an artificial collection of KPIs for each of those machines. The dataset related with WP5 will hold data about the coordinates of a machine within a virtual environment during the learning process.

The information provided can be summarized as:

- List of virtual machines
- List of machine virtual parameters
- List of machine KPIs
- Coordinates of cutting machine during learning process
- Learning parameters from RL process

5.2.1.6.2 Data openly available on Zenodo Platform

AIMEN and SCM will publish the described above data on ZENODO, KAGGLE or any other platform selected by the consortium

5.2.1.6.3 Data available only for the MAS4AI consortium partners

The MAS4AI consortium will have access to the data.

5.2.1.6.4 Data not openly available

In this chapter it is IMPORTANT to give a very good explanation, why the data cannot be shared openly.

5.2.2 How will data be accessed?

The data sets will be made available through the project website and other dissemination channels. No specific software tools will be needed to access the data. The anonymised data sets will be saved and stored in standardized formats such as xml, MS Word, pdf, csv or MS Excel to facilitate their exploitation (as well as guarantee their long-term accessibility).

5.3 How will data be made interoperable?

Interoperability will be ensured by the project partners by storing all data in an appropriate format that will make data accessible to all professionals who are interested in exploiting the data generated in the project.

The language used in the project will be the one in common use within research and manufacturing sectors, addressing the specific target audience of SMEs, larger companies and their ecosystems. Vocabulary will not constitute a barrier for data interoperability and reuse. Authors will only use words that are commonly understood by the sector of the target audience.

5.4 How will the data reuse be increased?

Data usage on side of third parties will be increased by providing anonymized data sets, depending on the use scenario (later in the project, training and test data for the agents). Furthermore, after the end of the project, relevant data generated that might be necessary to test and implement the created agents or other project results will be made available through

the project website. The consortium will maintain this project data in a reusable way for as long as possible after the end of the project. The duration is depending on the agreements among the project partners. Further data reuse will be achieved by linking dissemination activities with data collected in the project.

The project coordinator will be responsible for assuring the quality of the data by making sure that datasets follow the FAIR principles included in this document. The data quality assurance process will be in accordance with Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons regarding the processing of personal data and on the free movement of such data.

6 Allocation of resources

The costs to make the data FAIR compatible in the MAS4AI projects are covered in large parts by the grant of the project. This is especially applicable to the dissemination strategy as stated in 3 Open Dissemination Strategy.

7 Ethical aspects

In general, the MAS4AI project aims to share as much data as possible in relation to the project goals. This has the aim to make project results reproducible and verifiable by the scientific community and interested third parties. This is especially necessary in the recent wake of “replication crisis”, in which claims of research groups cannot be tested due to the lack of data [3].

However, there are also ethical aspects to be considered when sharing data, especially when it affects participants of the project or third parties. Data sharing can pose the risk that data can be associated with the identity, leading to harms such as stigmatization, discrimination and other harms in addition to the loss of privacy. Thus, all data will be checked under ethical and privacy viewpoints as stated in this document. In the MAS4AI project this will be especially necessary for agents or other work packages that relate to human-machine-interaction (HMI).

8 Public funding disclaimer

All data produced within the framework of the project will inform of the funding source by adding the following disclaimer and EU flag:

MAS4AI has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957204

Literature

[1] Wilkinson, M. D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

[2] File naming conventions. Available online: https://www.ed.ac.uk/records-management/guidance/records/practical-guidance/naming-conventions, last accessed: 2021-3-19.

[3] Meyer MN. Practical Tips for Ethical Data Sharing. Advances in Methods and Practices in Psychological Science. March 2018:131-144. doi:10.1177/2515245917747656.