EMERGE

WP9 Project Management and Coordination

D9.2 Data Management Plan

Version: 1.0

Date: 31/03/2023

UNIVERSITÀ DI PISA   TUDelft   University of BRISTOL

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN   DA VINCI LABS

Document control

Project title	Emergent awareness from minimal collectives
Project acronym	EMERGE
Call identifier	HORIZON-EIC-2021-PATHFINDERCHALLENGES-01-01
Grant agreement	101070918
Starting date	01/10/2022
Duration	48 months
Project URL	http://eic-emerge.eu
Work Package	WP9 Project Management and Coordination
Deliverable	D9.2 Data Management Plan
Contractual Delivery Date	M6
Actual Delivery Date	M6
Nature[1]	DMP
Dissemination level[2]	PU
Lead Beneficiary	UNIPI
Editor(s)	Davide Bacciu (UNIPI)
Contributor(s)	Claudio Gallicchio (UNIPI), Monica Marrucci (UNIPI), Ophelia Deroy (LMU), Nadine Meertnes (LMU), Sabine Hauert (UOB), Yasemin Turkylmaz-van der Velden (TUD)
Reviewer(s)	Renan Picoreti Nakahara (DVL)
Document description	This deliverable provides the first version of the Data Management Plan for the EMERGE project. It describes the data collected within the project as well as the plans for making those data, as well as other research outcomes, available and exploitable by the community.

---

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Version control

Version[3]	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.1	Monica Marrucci	14/02/2023	ToC based on Horizon Europe DMP Template
0.2	Davide Bacciu	24/02/2023	TOC revised and approved
0.3	Davide Bacciu	08/03/2023	First incomplete draft: partners contributions identified.
0.4	Davide Bacciu	13/03/2023	Added partner contributions. Description of FAIR data sections.
0.5	Davide Bacciu	30/03/2023	Complete draft ready for review
0.9	Renan Picoreti Nakahara	31/03/2023	Document approved by reviewer
1.0	Davide Bacciu	31/03/2023	Document released by Project Coordinator

[3] 0.1 – TOC proposed by editor; 0.2 – TOC approved by reviewer; 0.4 – Intermediate document proposed by editor; 0.5 – Intermediate document approved by reviewer; 0.8 – Document finished by editor; 0.85 – Document reviewed by reviewer; 0.9 – Document revised by editor; 0.98 – Document approved by reviewer; 1.0 – Document released by Project Coordinator.

EMERGE

Abstract

The report is the first version of the Data Management Plan for the EMERGE project. It describes the data collected within the project as well as the plans for making those data, as well as other research outcomes, available and exploitable by the community. The primary sharing platform for EMERGE's research outcome is identified and discussed.

Consortium

The EMERGE consortium members are listed below.

Organization	Short name	Country
Università di Pisa	UNIPI	IT
TU Delft	TUD	NL
University of Bristol	UOB	UK
Ludwig Maximilian University of Munich	LMU	DE
Da Vinci Labs	DVL	FR

EMERGE

Disclaimer

This document does not represent the opinion of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA), and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

This document may contain material, which is the copyright of certain EMERGE consortium parties, and may not be reproduced or copied without permission. All EMERGE consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the EMERGE consortium as a whole, nor a certain party of the EMERGE consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

Acknowledgement

This document is a deliverable of the EMERGE project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº 101070918.

Table of contents

List of abbreviations

CSV	Comma Separated Value (file)
DM	Data Manager
DMO	Data Management Officer
DMP	Data Management Plan
DoW	Description of Work
FACT	Fair, Accurate, Confidential, and Transparent
FAIR	Findable, Accessible, Interoperable, and Reusable
HLEG	High Level Expert Group
RI	Research Infrastructure
WP	Work package

List of figures

1. Executive summary

This document is a deliverable of the EMERGE project, funded under HORIZON-EIC-2021-PATH FINDER CHALLENGES-01-01 under grant agreement number 101070918.

The document structure is adapted from the Horizon Europe DMP template available here: https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf

The DMP describes the data management life cycle for the data to be collected, processed, and/or generated by EMERGE. As part of the effort of making research data findable, accessible, interoperable, and reusable (FAIR), this DMP includes information on:

- what data will be collected, processed and/or generated, and whether such data will be shared/made open access (Section 2);
- the handling of research data during & after the end of the project, which methodology & standards will be applied (Section 3);
- what further research outputs are produced and how they are made FAIR-ly available (Section 4);
- how data will be curated & preserved (Section 6 and 7).

The partners who will handle data in the EMERGE project are: UNIPI, TUD, LMU and the associated beneficiary UOB. DVL will not handle research data.

The deliverable "D9.2 Data Management Plan" describes the first version of the Data Management Plan (DMP) of the EMERGE project. The current deliverable is conceived as a living document, i.e. as needs arise and data or data management tools or standards change, we will include them in the report and we will submit the updated versions of the document, as foreseen by the Grant Agreement, by 31 March 2025 (D9.4 Updated Data Management Plan) and 30 September 2026 (D9.6 Final Data Management Plan).

To ease versioning and tracking of the changes, the table below reports the significant updates contained in the different versions of the DMP (and relevant documentation and resources) released to date.

Title	Description	Link
D9.2 Data Management Plan (*Date: 31 March 2023*)	First version of the DMP	*TBD upon deliverable approval by EC.*

2. Data summary

2.1 Type, format and size of collected/generated data

In the following we describe the original datasets that will be collected within the project, providing as much details as possible at the current stage of the project. Further details will

be added in future versions of this report. For each dataset we identify the partner responsible for its collection (in bold) and we list all partners which are planning to handle the data therein.

Moral Psychology & Experimental Ethics Data (**LMU**; UNIPI, UOB)

TYPE

This data will be collected within the scope of the activities of WP2 (in particular T2.2 and T2.3) aimed at testing how the new concept of collaborative awareness stands regarding two key ethical risks for humans when relating to AI agents, i.e. the lack of understanding and the tendency to anthropomorphise their awareness. The leading partner managing the experiments and data collection is LMU.

For the purposes above, we will collect questionnaires, demographic data, and behavioural data in psychological experiments. This will be done both via online experiments as well as physical lab-based experiments.

We will conduct empirical studies to investigate human interactions with artificial agents using established and fairly well-known behavioural game theory methods. The collected data would involve choice in various scenarios, judgements of situations, and ratings of preferences and emotions. This will entail recording human participants' decisions and other performance metrics in the given tasks, and collecting demographic data such as the participant's age, gender, nationality.

FORMAT

At the current stage, data will likely be collected under the form of CSV files. All data will be stored in offline digital storage. For online experiments data will be collected via platforms such as Mturk and Prolific: these platforms maintain the identification and only an encrypted ID will be entered into the offline storage at the partner site. For in-laboratory based experiments, the data is collected on hardcopies, and again only an encrypted ID will be stored on the digital storage.

SIZE

We estimate that the data size will be less than 100 GB. The sample size of each experiment will generally be around 100 participants. As part of the EMERGE activities, we expect to perform about 10 experiments, with a total sample size ranging in about 1000-1500 people.

Swarm Experimental Data (**UOB**; UNIPI, TUD)

TYPE

This data will be collected within the scope of the activities of WP5 and WP6, and will comprise logs of physical experiments with the DOTS robots. This data will serve for demonstrating the Awareness in Swarm use case (T6.3) and for the validation of the emergent awareness measurement and control methodologies from T5.1-T5.3. The leading partner managing the experiments and data collection is UOB.

FORMAT

Data will comprise logs of onboard sensors and state information from the DOTS robots, together with video files of the experiments. These will be complemented by scripts/code for experimental analysis. These data and associated resources will be made openly available according to the provisions in Section 3.

SIZE

Expected size is Gigabytes per experiment.

<div align="center">Swarm user studies (UOB; UNIPI, TUD, LMU)</div>

TYPE

This data will be collected within the scope of the activities of WP5, and in particular to assess the ability of users to monitor and control emergent awareness of our robot swarm (activities T5.1, T5.3 and T5.4). The leading partner managing the experiments and data collection is UOB.

For this purpose, we will collect anonymised questionnaires and interviews, associated with logs of experiments. The data will also be linked to scripts for its analysis.

These data and associated resources will be made openly available in anonymised form, according to the provisions in Section 3.

FORMAT

At the current stage, questionnaire and experimental data will likely be collected under the form of CSV files. Interviews will be collected as blurred videos.

SIZE

Expected size is Gigabytes per experiment.

<div align="center">Simulated Robotic Traces (UNIPI; TUD)</div>

TYPE

Data will comprise simulated robotic traces, generated using available software simulators for the robotic platforms used in EMERGE (e.g., DOTS, Kilobots swarms, soft robotic components). This data will be collected mainly for the purpose of validating the archetype computing system and the associated learning functionalities, developed jointly by WP3-WP4.

Data will be made openly available according to the provisions in Section 3.

FORMAT

CSV files.

SIZE

Expected size is about tens of Megabytes per sample.

2.2 Reuse of existing data and its origin

In the first phases of development of the Archetype Computing System, we will make use of publicly available datasets to benchmark in learning architectures and algorithms, also for publication purposes. Examples of datasets currently used to benchmark learning models are given in the following.

MNIST – A public dataset representing a classic benchmark for neural networks. The dataset contains 70000 28x28 images representing handwritten digits. This dataset is used to test different architectures and learning strategies, in different settings. (SOURCE: http://yann.lecun.com/exdb/mnist/)

CIFAR10 - A public dataset representing a classic benchmark for neural networks. This is a dataset of 50,000 32x32 colour training images and 10,000 test images, labelled over 10 categories. As for the MNIST dataset, it is used to test different architectures and learning strategies, in a variety of settings. (SOURCE: https://www.cs.toronto.edu/~kriz/cifar.html)

WESAD – A publicly available dataset for recognition of human emotional state recognition. The dataset features multimodal physiological data gathered through wearable devices from 15 subjects. (SOURCE: https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/)

3. FAIR data

3.1 Making data findable, including provisions for metadata

The data collected/generated within the project will be made available to all the project partners for its duration. The datasets collected as part of the project (referenced in Section 2.1) will also be made openly available to the community, in anonymised form whenever personal data is involved.

All data released by the project will be made available free of charge to the community through services integrated in the European Open Science Cloud (EOSC). The primary sharing service of choice for the implementation of EMERGE FAIR policies is the SoBigData (http://www.sobigdata.eu) catalogue. SoBigData is a European open science Research Infrastructure (RI) developed through a series of EU funded initiatives, H2020 SoBigData (G.A. n. 654024) and H2020 SoBigData++ (G.A. n. 871042), coordinated by the Italian National Research Council and heavily contributed by UNIPI, through members of the EMERGE consortium.

SoBigData provides a broad set of services, data, methods, training materials and technologies for Artificial Intelligence and Data Science within operative guidelines including FAIR principles, EU Ethical, Legal, Social, Economic, and Cultural (ELSEC) values, and FACT (Fair, Accurate, Confidential, and Transparent) principles are key aspects in the research infrastructure development. The choice of relaying on this platform as the reference one to implement EMERGE FAIR policy is motivated by:

1. The integration within the EOSC and the possibility of defining in the SoBigData Catalogue an organisation specific for EMERGE, to promote the visibility of the project and to appropriately brand its research outcomes.
2. The broad scope of the RI which allows sharing of data, but also code and legal/ethical workflows, which matches the nature of the research outcomes that will be FAIRly shared by EMERGE.
3. The strong focus of the SoBigData RI on social, ethical and cultural aspects, which is very much in line with the philosophy of EMERGE, which blends scientific advancement, technological innovation and ethical/philosophical frameworks.
4. The simplified interaction with the RI managers enabled by the involvement of UNIPI in the SoBigData initiatives.

Datasets (as well as other types of resources) are made searchable and accessible in the SoBigData RI through the SoBigData Catalogue web-service. A user can search and explore the set of resources available by inserting a set of keywords, obtaining a ranked list of matching results. The user can also filter and organise results based on a taxonomy of resource properties, including the type (dataset, method, application, training material, experiment), the format (PDF, CSV, JSON, …), the licence, etc. Figure 1 shows a snapshot of the homepage of the SoBigData Catalogue, along with some example organisations included (which will be extended, for instance, with the inclusion of EMERGE).

EMERGE data will be complemented with metadata needed for SoBigData indexing, which are detailed in Appendix A. This metadata will include, among others, information detailing the acquisition process, the origin of the data, information on known anomalies, intended use and licensing. Whenever possible/meaningful the data will be complemented with performance results from baseline methods (e.g. from AI models developed in the EMERGE). To foster reproducibility and simplify data use, we will also release and associate source codes for replicating the baseline results.

In order to ensure the most effective diffusion of the data, datasets may also be released in other sharing platforms, for instance taking into consideration aspects such as:
1. the interests of the community revolving around the platform, its popularity, adequacy of the access interface to the specificity of the data.
2. availability of a scientific publication associated with the dataset, if the publisher offers an associated data sharing platform compliant with FAIR and FACT policies.
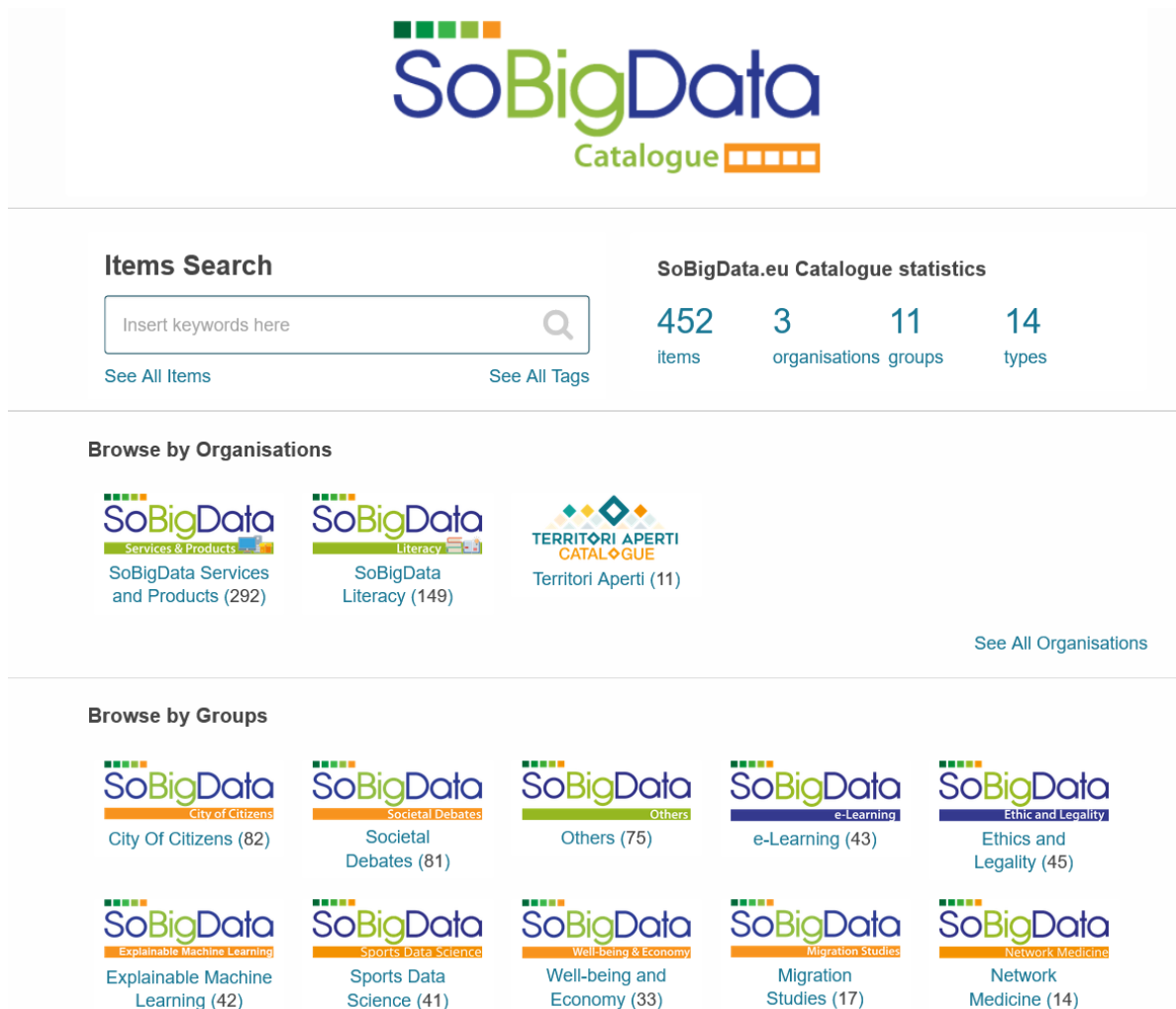
12

Figure 1 - SoBigData Catalogue Home Page - March 2023

3.2 Making data accessible

SoBigData defines two main ways to provide data access to the RI, related to the availability of the offered resource: on-line by e-infrastructure facilities, on-site by visiting the institution that is the data controller of the dataset. EMERGE is expected to deliver its resources through the online facility only. Access to every resource in SoBigData RI requires the user to be registered and logged. This enables the platform to keep track and statistics of the accessed datasets, which will be leveraged to assess the usage and impact of the EMERGE research outcomes.

SoBigData metadata have an ethical and legal section that defines why the user cannot freely download a dataset or if a dataset can be accessed, e.g., considering the geographical restriction of using a particular dataset. At the current stage, we do not foresee making use of such type of access restrictions in EMERGE.

Every dataset added to the SoBigData RI is subject to an internal review of the SoBigData Ethical and Legal board before being made publicly available. We expect this scrutiny to be

relevant for the EMERGE datasets dealing with moral psychology experiments and swarm users experiments.
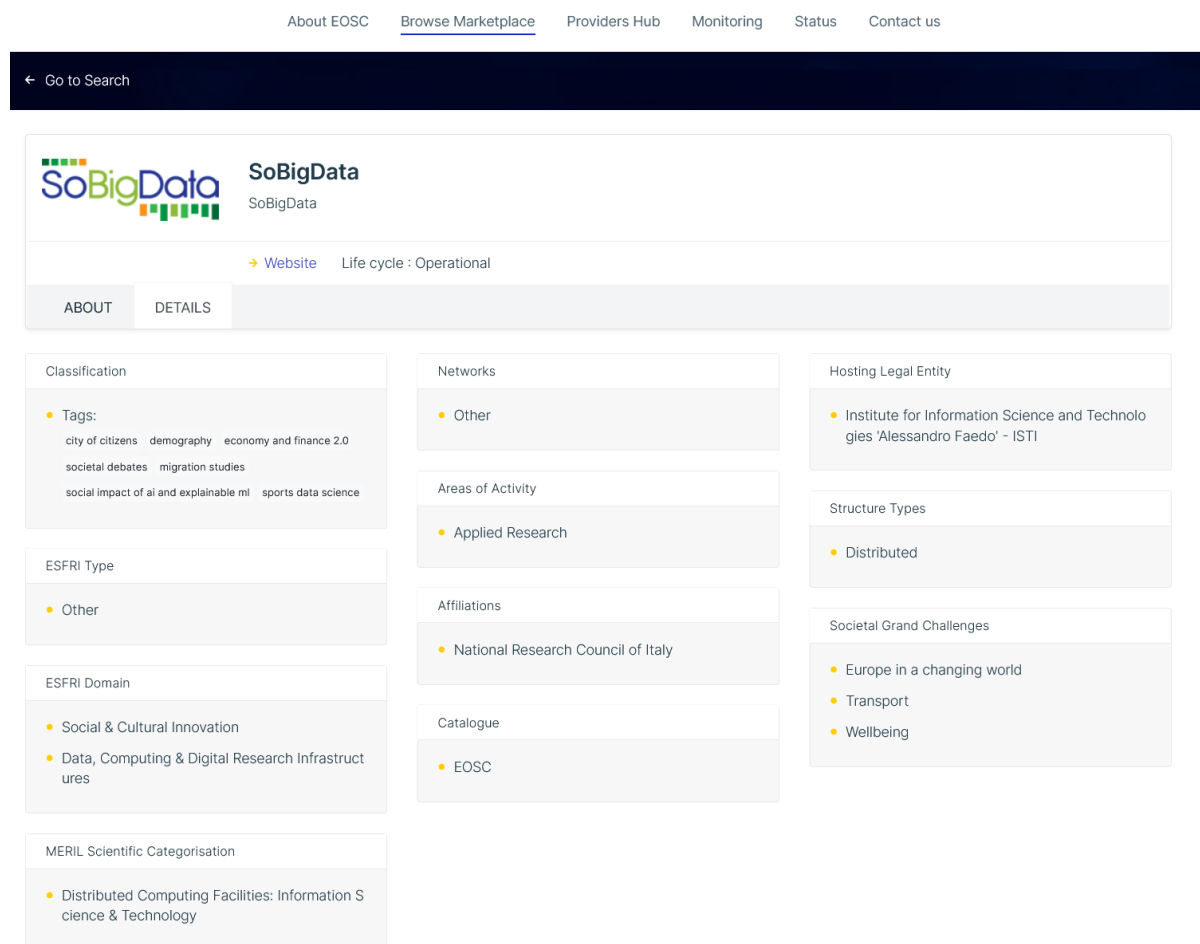
3.3   Making data interoperable

Figure 2 - SoBigData entry as Provider in the EOSC Marketplace (https://marketplace.eosc-portal.eu/providers/eosc.sobigdata).

Released data will be in standard formats widely adopted by the Artificial Intelligence community, such as formatted text readable by standard Python Pandas API methods, allowing easy code-based access, use and combination with other data sources. As a general principle data will be openly shared on SoBigData as soon as possible. This might involve short embargo periods to give time to associated publications to be accepted (typically project, following the determinations in the project Contractual Agreement.

SoBigData is integrated as a service provider in the EOSC Marketplace (see Figure 2 above). In addition to that, the SoBigData e-infrastructure defines a Workspace service to make data interoperable and employable by other services, e.g. JupyterHub. The workspace is an online virtual research environment, supporting secure and controlled data storage and sharing. Each virtual research environment has an associated workspace, partitioned in private and

public areas, where users can store, access, and share documents and results related to their activities. The SoBigData RI allows to select and execute available methods on the datasets: this is performed within the workspace of the user and can be made publicly available.

3.4 Increase data re-use

We expect most of EMERGE project data to be re-usable for a long time, given the simulation-based nature of several datasets. The choice of a consolidated sharing platform such as SoBigData is also intended to ensure persistence of shared data across time. Datasets in SoBigData include a unique reference and an assessment of their nature, scale, and availability (such as related scientific publications, privacy issues, data governance policies, licensing, or similar resources). The preservation and re-using procedures bundled in the platform describe how data is stored, which technology is used, and for how long the data is available.

EMERGE data will be made available together with accompanying code that will allow, for instance, to run pre-processing and data cleaning, to replicate experiments with benchmarking models and EMERGE methodologies, and to generate further data (in case of simulated datasets). The SoBigData infrastructure will be a key enabler for such an increased data reuse, as it provides data, methods and computational infrastructures made accessible to the user via an integrated interface through the Workspace service.

4. Other research outputs

The research outcomes that EMERGE will make available to the community are not limited to data alone.

One of the key outputs of WP2 will be the release of an ethical toolkit including both templates and codes to generate (a) text and graphic vignettes, (b) game theoretic scenarios, (c) surveys, (d) visualisation of dimensions of awareness for different systems, and (d) visualisation of distribution and simple statistics regarding ethical responses. The ethical toolkit will therefore be a mixture of workflows, code and training material whose nature matches   very well the resource types in the SoBigData RI. Additionally, the focus of the SoBigData platform and its community is heavily oriented towards social, ethical and regulatory implications of AI. We expect that such matching interests will promote the impact and favour the uptake of the EMERGE ethical toolkit.

A main research output for the technical WPs in EMERGE will be the release of open software, including all software and codes related to methodologies presented in academic papers, robotic simulators, and all the code needed to replicate EMERGE published experiments. Additionally, UOB will follow an open-hardware approach for robots and interfaces developed as part of EMERGE.

The released software and code will be shared through the SoBigData interface and, possibly, in additional tools chosen by the partners responsible for the resource, namely Bitbucket for UOB and GitHub for UNIPI/TUD.

5. Allocation of resources

The costs for data acquisition and generation are covered by the project budget. The data sharing platforms identified in the previous sections are free of charge, hence there is no need to plan additional costs for long term preservation (e.g. following the end of the project). Even when considering data repositories associated with scientific publishers, these are typically free of charge (costs for open publication fees are already foreseen in the project budget).

The Data Management Officer (DMO) for EMERGE is the Coordinator, Davide Bacciu (davide.bacciu@unipi.it). The DMO is responsible for managing application of the data sharing policies and for supporting the partner in charge for the data repository in their implementations. Leading partners for each dataset have been identified in Section 2.1. These partners are primarily responsible for data collection/generation, its secure storage and preservation, for providing access to it to each project partner and, ultimately, for preparing data for sharing on public repositories. Each partner has identified a Data Manager (DM) who will implement the data sharing procedures locally. The list of the DMs along with their contact is provided in the following:

- UNIPI: Davide Bacciu (davide.bacciu@unipi.it)
- TUD: Yasemin Türkyilmaz-van der Velden (datasteward-3mE@tudelft.nl)
- LMU: Nadine Meertens (nadine.meertens@lrz.uni-muenchen.de)
- DVL: Renan Picoreti Nakahara (renan.picoreti@davincilabs.eu)
- UOB: Simon Jones (simon2.jones@bristol.ac.uk)

6. Data security

For the openly shared resources, the SoBigData RI ensures privacy and data integrity in any connection between a client (e.g., a web browser) and a SoBigData e-infrastructure server. Connections are characterised by (i) symmetric cryptography to encrypt the data transmitted, (ii) identification of the communication parties using public-key cryptography, and (iii) message integrity checks.

Project private information (not meant for sharing) collected as part of the moral psychology and swarm user studies will be maintained either on hard-drives (disconnected from the network), or on a server in the local area network of the responsible partner, in encrypted folders or within a Database Management System with an access control management. Access to these data will be granted upon agreement with the DPO and the data maintainer partner. Further details on personal data protection measures are provided in the dedicated "D2.1 Ethics Monitoring" deliverable.

7. Ethics

Project data will be used to train and validate machine learning models. As such both data and the trained machine learning models will need to adhere to the ethical guidelines of the HLEG on Artificial Intelligence. These aspects are discussed and articulated in a dedicated

deliverable D2.1, as indicated in the Description of the Action document. All datasets in the project are either synthetically generated or collected in experiments involving only volunteers. In the latter case, a detailed statement of consent is collected, and data is duly anonymised.

8. Other issues

We will primarily follow the EU General Data Protection Regulation (GDPR). Additional procedures and provisions for the specific partners are detailed below.

LMU: The behavioural and experimental ethics studies performed by LMU will be pre-evaluated and approved by the ethics committee of either the Faculty of Philosophy, Philosophy of Science and Religious Studies, or the Psychology faculty at Ludwig-Maximilians-Universität München (LMU Munich).

TUD: Activities will be carried in compliance with the TU Delft Research Data Framework Policy stating that research data, code and any other materials needed to reproduce research findings are appropriately documented and shared in a research data repository in accordance with the FAIR principles (Findable, Accessible, Interoperable and Reusable) for at least 10 years from the end of the research project, unless there are valid reasons not to do so.

UOB: The University of Bristol provides a Research Data Service as part of the Library and supports pan-University research activities to help realise the impact of research on society. They help individual researchers or major interdisciplinary research initiatives to plan, manage and share research data. More information can be found here: http://www.bristol.ac.uk/staff/researchers/data/.

DVL: The project may maintain a mailing list of email addresses (and, in some cases, names) of people participating in the events organised by the Consortium. This mailing list will be collected upon registration to the event, provided the registrant authorises the inclusion in the list. This mailing list will be held on the web-based email marketing service GetResponse. The mailing list is administered by and under the responsibility of DVL. Project members cannot view any details about subscribers. The subscriber list contains personally identifiable information (e-mail addresses and potential names) and is subject to GDPR. A privacy statement will be accessible, included (or linked) in each message posted to the list. The list will be used only to send communications about upcoming events organised by the consortium.

Appendix A. SoBigData Metadata

The following table summarises the Metadata available for managing the SoBigData catalogue. Mandatory entries are marked with an asterisk. The standard currently in use by SoBigData is Datacite (https://schema.datacite.org/)

Field name	Standard	Description
Title*	DataCite:Title	
Description	DataCite:Description @type="abstract"	How the data was produced: describe process/algorithm used, intended use, how the data can be reused: describe format, where to find schema/model, Issues with the content nature and reuse: describe dynamic constraints (no IPRs), exceptions, etc.
Tag*	DataCite:Subject	Defined a set of pre-defined tags. More than one tag can be selected for each dataset.
Licence		Report a set of pre-defined licences. Licence definitions and additional information can be found at opendefinition.org
Selected Licence Url		Report the link to the licence selected on the previous field. generated automatically based on the selected licence.
Visibility		Two values admissible (public/admissible). Restricted items can only be accessed by certain users, while public items can be accessed by anyone.
Author*		The name of the author of the item into the catalogue. This is not necessarily the creator of the dataset.
Author Email*		The mail address of the author of the item into the catalogue.
Types	DataCite:Subject	This field defines the type of the item we are inserting into the catalogue. Select "dataset" for adding a dataset.
Item Groups	DataCite:Subject	This field is required for technological reasons (this field will be removed).
Group	DataCite:Subject	Enable the selection of the SoBigData group the item belongs to.
Thematic Cluster*	DataCite:Subject	The list of the thematic clusters defined by the H2020 SoBigData++ project.
External Identifier	DataCite:Identifier	This field applies only to datasets that have already been published. Insert here a DOI, a handle, and any other Identifier assigned when publishing the dataset elsewhere.
Creator*	DataCite:Creators	The name of the creator of the dataset, with email and ORCID. The format should be: family name, given name, email, ORCID. Example: Smith, John, js@acme.org, orcid.org/0000-0002-1825-0097

CreationDate*	DataCite:Date @type = "creation"	The date of creation of the dataset (different from the date of registration of the dataset automatically added by the system).
Area	DataCite:Subject	Defines a Sub-community specific area.
TimeCoverage	DublinCore:coverage .temporal using RKMS-ISO8601	Defines a list of time intervals related to the dataset.
Spatial		Defines the Geo Location of the data available in the dataset. The value must be a valid GeoJSON geometry. This field is specifically used for mobility data.
ProcessingDegree*		Primary (raw data) or secondary (processed data) dataset.
ManifestationType*		Virtual (accessible in streaming from remote sites), replica (copy of data in remote sites, e.g. DBPL), original (collection of data produced and kept in local infra by data provider).
Language	DataCite:Language	The primary language of the dataset (using ISO 639-3). You can see ISO 639 Code Tables here: http://www-01.sil.org/iso639-3/codes.asp
RelatedPaper		Reports a complete reference to an associated work.
Accessibility*		Defines how the access to the resource is regulated: Virtual Access or Transnational Access.
Availability*		Defines how the resource is offered by the RI.
Size	DataCite:size	This field defines the size of the dataset in your domain/mind, not disk size.
DiskSize	disk size in Mbytes	
Format	DataCite:Format	Defines MIME or extension.
FormatSchema		Link to Schema of the dataset.
IP & Ethical Fields (incl. Data-protection Metadata)		
IP & Ethical Fields		
PersonalData*		Reports if the dataset contains personal data
PersonalSensitive Data		Reports if the dataset contains personal sensitive data
ChildrenData*		Reports if the dataset contains children data.
Consent of the data subject*		Consent of the data subject. Data subject signifies her/his agreement to personal data relating to him being processed
Consent obtained also covers the envisaged transfer of the personal data outside the EU*		Consent obtained also covers the envisaged transfer of the personal data outside the EU.
DataProtectionDirect ive*		Report the law or protocol number and the institution related to Data Protection.

IP/Copyrights	DataCite:rights	Whether the dataset is covered by any rights: copyright, related rights, database right, know-how, proprietary, etc.
Field/Scope of use*		List of alternatives.
Basic rights*		List of alternatives.
Restrictions on use		Any restrictions on how/where the dataset may be used.
Sublicense rights*		Restrictions on/availability of sublicensing rights
Attribution requirements		The text exporting how the user needs to acknowledge the source when using/distributing data/developing service.
Display requirements		Whether the user, when displaying the dataset in any media or form, must follow certain display requirements, e.g. attach copyright notice.
Distribution requirements		Whether the user, when distributing the dataset, if allowed, must follow certain requirements.
Territory of use*		Defines in what territory the dataset may be used.
Licence term		If reported this field reports the period of time during which the dataset may be used.
Requirement of non-disclosure (confidentiality mark)		Whether the dataset bears confidentiality marks/may be used and shared subject to the obligation of non-disclosure.
Custom Field(s):		Custom fields are customizable metadata that will be added to the item. You can remove them at any time until you create the item.