



D1.18 “Data Management Plan – Initial Version”

Author:	ETA
Work Package:	WP1 - Beyond COORDINATION: AI DIH consortium from management to governance
Delivery date:	25.04.2021
Due date:	31.03.2021
Classification:	PU

The AI REGIO Project owns the copyright of this document (in accordance with the terms described in the Consortium Agreement), which is supplied confidentially and must not be used for any purpose other than that for which it is supplied. It must not be reproduced either wholly or partially, copied or transmitted to any person without the authorization of the Consortium.



H2020 Innovation Action - This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N. 952003



Status of deliverable

Action/role	Name	Date
Submitted by	ETA	23.04.2021
Responsible (WP leader)	POLIMI	25.04.2021
Approved by (internal reviewer)	Francisco Souza (SKU)	16.04.2021

Date	Revision version	Author	Comments
16.02.2021	v0.1	Marina Da Bormida (ETA)	Table of Content – Structure of the document and assignment of writing responsibilities
16.03.2021	v0.2	Marina Da Bormida (ETA), Sergio Gusmeroli (POLIMI), Angelo Marguglio (ENG), Sarah Mortimer (SEZ)	First round contributions
19.03.2021	v0.3	Marina Da Bormida (ETA)	First draft of the document
29.03.2021	v0.4	Nenad Stojanovic (NISSA), Saadia Dhouib (CEA)	Second round contributions
02.04.2021	v0.5	Marina Da Bormida (ETA)	Second draft of the document
09.04.2021	v0.6	Marina Da Bormida (ETA), Claudio Palasciano (POLIMI), Silvia Razzetti (POLIMI), Nenad Stojanovic (NISSA)	Third round contributions
12.04.2021	v0.7	Marina Da Bormida (ETA)	Document ready for Peer Review and Quality Check
16.04.2021	v.0.8	Francisco Souza (SKU)	Peer Review and Quality Check
23.04.2021	v1.0	Marina Da Bormida	WP1 leader Quality Check



		(ETA), Sergio Gusmeroli (POLIMI), Claudio Palasciano (POLIMI)	
25.04.2021	v1.1	Marina Da Bormida (ETA), Claudio Palasciano (POLIMI)	Official release and submission to the EC

Name	Organisation	E-mail
Marina Da Bormida	ETA	marina.cugurra@eta-one.com
Sergio Gusmeroli	POLIMI	sergio.gusmeroli@polimi.it
Claudio Palasciano	POLIMI	claudio.palasciano@polimi.it
Silvia Razzetti	POLIMI	silvia.razzetti@polimi.it
Angelo Marguglio	ENG	angelo.marguglio@eng.it
Nenad Stojanovic	NISSA	nenad.stojanovic@nissatech.com
Sarah Mortimer	SEZ	Sarah.Mortimer@steinbeis-europa.de
Saadia Dhouib	CEA	saadia.dhouib@cea.fr



Contents

1	Executive Summary.....	6
2	Introduction	7
2.1	About this deliverable.....	7
2.2	Document structure.....	7
	The is structured as follows:.....	7
3	AI REGIO participation to the Open Research Data Pilot	8
4	Data Summary.....	9
4.1	AI REGIO Research and Innovation Data.....	9
4.1.1	Data related to the Experiments	9
4.1.2	Data related to AI REGIO DIHs	12
4.2	Scientific publications Data.....	16
4.2.1	Publishing infrastructure for Open Access	16
4.3	AI REGIO Deliverables Data	18
5	Making AI REGIO DATA Findable Accessible Interoperable and Re-usable (FAIR)	18
5.1	Making data findable.....	19
5.2	Naming convention	19
5.2.1	Search keywords.....	19
5.2.2	Versioning	20
5.3	Making data openly accessible.....	20
5.4	Making data interoperable	20
5.5	Making data reusable.....	20
6	Data Security	21
7	Storage and Preservation of Data	22
7.1	Storage.....	22
7.1.1	Backup and Recovery.....	23
7.1.2	Data Archiving.....	23
7.2	Preservation	23
8	Data Quality Assurance.....	24
9	Ethical Aspects and Legal Compliance.....	24
10	Allocation of resources.....	25
11	AI REGIO Datasets.....	25
11.1	Datasets categories	26
11.2	Update of AI REGIO Dataset templates.....	26
11.2.1	General template.....	26
11.2.2	Technical template.....	27
12	References.....	28



Tables

TABLE 1: AI REGIO EXPERIMENTS	9
TABLE 2: AI REGIO EXPERIMENTS DATA SUMMARY	10
TABLE 3: DATA SUMMARY FOR DIHS SERVICE PORTFOLIO	13
TABLE 4: DATA SUMMARY FOR DIHS ORGANIZATION DETAILS	14
TABLE 5: DATA SUMMARY FOR DIH CUSTOMER PIPELINES	14
TABLE 6: DATA SUMMARY FOR DIHS BUSINESS MODELS	15
TABLE 7: AI REGIO PILOT DATA MANAGERS	25
TABLE 8: GENERAL TEMPLATE	26
TABLE 9: TECHNICAL TEMPLATE	27

Figures

FIGURE 1 - OPEN RESEARCH EUROPE PUBLISHING PROCESS	17
--	----

Abbreviations and Acronyms:	
AAS	Asset Administration Shell
AI	Artificial Intelligence
APC	Article Processing Charges
API	Application Programming Interface
AutomationML	Automation Markup Language
B2MML	Business To Manufacturing Markup Language
CA	Consortium Agreement
CI	Collaborative Intelligence
DIH	Digital Innovation Hub
DMP	Data Management Plan
DOA	Description of Actions
DPO	Data Protection Officer
EC	European Commission
ECTL	European Centre for Living Technology
ERP	Enterprise Resource Planning
ES	Ethical Strategy
FAIR	Findable, accessible, interoperable and reusable
FMCEA	Failure Mode, Effects and Criticality Analysis
FTP	File Transfer Protocol
GA	Grant Agreement
GDPR	General Data Protection Regulation (Regulation EU 2016/679)



HDF	Hierarchical Data Format
HF	Human Factor(s)
HMI	Human-Machine Interface
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LPKF	Leiterplatten-Kopierfrasen (Circuit Board Copy Milling)
MES	Manufacturing Execution System
OA	Open Access
OPC - UA	Open Platform Communications (OPC) - UA (Unified Architecture)
ORDP	Open Research Data Pilot
ORE	Open Research Europe
OWL	Ontology Web Language
OSAI	Observatory on Society and Artificial Intelligence
PCC	Project Coordination Committee
PDM	Pilot Data Manager
RAMI	Reference Architectural Model Industrie
RFC	Request for Comments
ROS	Robot Operating System
SCADA	Supervisory Control And Data Acquisition
SME	Small Medium Enterprise
SSH	Secure SHell
SSL	Secure Sockets Layer
TC	Technical Coordinator
UI	User Interface
WP	Work Package
XML	eXtensible Markup Language



1 Executive Summary

This deliverable describes the initial version of the AI REGIO Data Management Plan, in accordance with the strategy for its elaboration, as depicted in D1.11 “Ethics Assessment and Data Management Strategy”.

The DMP has been developed in compliance with the guidelines provided by the Open Research Data Pilot initiative and its content rotates around the FAIR principles and OA paradigm. This is reflected first of all in the Data Summary, where in-depth information on AI REGIO Research and Innovation Data are provided, with a special focus on:

- initial data related to the experiments, lingering over their cluster (“Product Engineering and Lifecycle management”, “Factory Efficient and Sustainable Manufacturing”, “Quality Control and Predictive Maintenance”, “Robotics and Human Interaction”)
- initial data related to DIHs, where data are classified in four different types of data to be collected: Service portfolio, Organisation details, Customer pipeline, Business model.

For each of these clusters and DIHs’ data typologies the information compliant with the ORDP Data Summary guidelines have been collected in a table format, in order to make more evident the key peculiarities.

As for the scientific publications and the data needed to validate the results presented in them, the AI REGIO Consortium plans to adhere to the ORDP principle “as open as possible, as closed as necessary”, in compliance with Article 29.2 of the Grant Agreement. The Open Research Europe scientific publishing service will be explored for rapid and transparent publishing and to facilitate open, constructive research discussion, whilst the repository for AI REGIO open access publications will be based on the free Zenodo platform (zenodo.org). The AI REGIO “community” within Zenodo will be the common area where to upload scientific publications.

The methodology and steps for adhering to the FAIR approach are depicted in a dedicated section (Sect. 5), which provides all the details on how the Consortium will make data findable, accessible, and interoperable and reusable, as well as will ensure their preservation and open access, with a view to increasing its re-use. This also includes AI REGIO naming conventions.

Great attention is also paid to the Security issues. Measures and choices have been taken to securely handle with any collected/generated data throughout its entire lifecycle, in order to safeguard them against accidental loss and/or unauthorised manipulation. Likewise, insights on how to ensure the high-level quality of collected and generated data, as well as their the storage and preservation are addressed. These themes are relevant considering the possibility for the AI REGIO project to manage high volumes of data and also in view of foreseeing situations of exponential growth of the data volume or data loss. At the same time, ethical and legal issues are taken into account by the Consortium, which dedicated a consistent part of D1.11 to depict the strategy and road-map for addressing them.

The deliverable contains information on the allocation of resources and responsibilities for making data FAIR- compatible and data management, including the list of AI REGIO Experiment Data Managers.

Furthermore, the datasets to be produced or collected in AI REGIO will be classified into categories on the basis on the existing clustering of experiments, whilst the Dataset templates (both “General” and “Technical”) are updated. As the nature and extent of the datasets can evolve during the project, more detailed descriptions will be provided in the next version of the DMP, in conjunction with the datasets list.



The DMP is a living document within AI REGIO: its next release will be outlined in D1.12 “Ethics Assessment and Data Management Plan” (at Month 18), whilst the final version will be released at the end of the project in D1.19 “Data Management Plan – Final Version” (at Month 36).

2 Introduction

2.1 About this deliverable

This deliverable describes the initial version of AI REGIO Data Management Plan, which has been developed in accordance to the methodology and preliminary remarks reported in D1.11.

The Data Management Plan is strictly aligned with FAIR principles and OA paradigm, in order to adhere to the Open Research Data Pilot in a consistent and effective manner.

2.2 Document structure

The is structured as follows:

- **Section 3** offers an overall snapshot of AI REGIO project's **participation to the Open Research Data Pilot**, including key remarks on our approach and commitment in this regards;
- **Section 4** describes the **Data Summary**, providing in-depth information on AI REGIO Research and Innovation Data, lingering on data both related to the experiments and to DIHs, besides on data pertaining to scientific publications;
- **Section 5** deepens how we intend to adhere to the **FAIR approach**, focusing on how to make data findable, openly accessible, interoperable and reusable, as well as AI REGIO naming conventions;
- **Section 6** refers to the **securely handling of any collected/generated data** throughout its entire lifecycle, in view of safeguarding it against accidental loss and/or unauthorised manipulation;
- **Section 7** depicts the **plan for storage and preservation of data**, which is key given the possibility of the AI REGIO project to manage high volumes of data and also in view of foreseeing situations of exponential growth of the data volume or data loss;
- Section 8 contains insights on how we will ensure the high-level quality of collected and generated data;
- **Section 9** refers to D1.11 in relation to **legal and ethical issues**, considering that the strategy and road-map for addressing them have already been deeply described in such deliverable;
- **Section 10** reports how we foresee to **allocate resources** for data management according to the ORDP and list the **AI REGIO Experiment Data Managers**;
- **Section 11** provides an overview of the different types of datasets to be produced or collected in AI REGIO (identified at this stage of the project), notably grouping them in **data categories**. It also provides an update of **AI REGIO Dataset templates**, both general and technical, which were inserted in D1.11.
- In **Section 12 conclusions** are drawn.



3 AI REGIO participation to the Open Research Data Pilot

Since December 2013, the European Commission is committed to open data through the Pilot on Open Research Data, as part of the Horizon 2020 Programme. The initiative is directed to “improve and maximize access to and re-use of research data generated by projects for the benefit of society and the economy”. In its frame, results generated from publicly-funded research should be disseminated more broadly and faster, for the benefit of researchers, citizens and innovative industry.

Open Access Policy allows on the one hand to accelerate the discovery process and ease those research results to reach the market (thereby meaning a return of public investment) and, on the other had, to avoid a duplication of research efforts and therefore to a better use of public resources. Such Policy also increases the visibility and scientific impact of the performed research, which is beneficial for the researchers themselves: it might imply a significantly higher number of citations, as well as an improvement in the collaboration potential with other institutions/organization within new projects, among others.

AI REGIO is fully committed in meeting OA requirements and indications, both as regards scientific publications and as regards the research data needed to validate the outcomes presented in the scientific publications ("underlying data"). For this reason, in order to effectively supply these data, we have started since the beginning of the project how we are going to manage, store and share the data (including make them public) we are about to create, generate and/or use. In line with this choice, the strategy functional to the development of the Data Management Plan (DMP) was conceived and reported in D1.11.

The effective participation of AI REGIO to the Open Research Data Pilot initiative is therefore ensured thanks to our choices in relation to:

- Data availability: Project data will be made publicly available as much as possible, within the boundaries set by the Consortium Agreement and other relevant pieces of contractual and/or legal framework relevant to AI REGIO. For this reason, it is also planned that different access levels will be granted for different types of data, as well as personal data will be regulated by data protection rules and obscured. Notes and recordings and from meetings and workshops as well as survey results will be anonymized. Anonymized data will be available in open-access mode.
- Scientific Publications, for which Open Access will be ensured as much as possible: they will be open, unless special requirements, constraints or policies will force to non-open publications.
- Research Data will be made accessible to the research and professional community to the maximum extent, exploiting research data archiving and availability of tools like Zenodo digital repository.

Further details and insights on how we are committed to participate to the Open Research Data Pilot are provided by the following sections.



4 Data Summary

4.1 AI REGIO Research and Innovation Data

4.1.1 Data related to the Experiments

AI REGIO Research and Innovation Data comprise first of all data related to the experiments of the project, which will be conveyed to the AI REGIO AI4Manufacturing Toolkit and Data4AI integrated platforms in order to be accessible and “executable” by a worldwide users’ constituency. The following table summarizes AI REGIO Experiments, including the partners in charge of each of them.

Table 1: AI REGIO Experiments

Experiment. Number	Experiment Name	Responsible partner
EXP-01	Machine Vision for Warehouse Optimization	GUALINI
EXP-02	NLP for Maintenance Report Analysis in Machine Tools	IMECH
EXP-03	Anti-Tampering Devices for Connected Objects	S2P
EXP-04	AI for better life cycle and project management for plastronic product	SWARM
EXP-05	AI-based Predictive Dynamic Production Planner	HOHNER
EXP-06	AI-Supported Robot Trajectory Optimization	KAUTENBURGER
EXP-07	Predictive Analytics for Autonomous Mobile Robots	ARCULUS
EXP-08	AI-enabled Operators Assistance through AR	BRAINPORTDev
EXP-09	AI-based Process Control c/o S/PARK DEVENTER	SKU
EXP-10	AI-based Process Control c/o ARMAC	SKU
EXP-11	Automatic Capability Matchmaking for Re-configurable Robotics Platform	TAU
EXP-12	IoT and AI for Real-time Production Planning Adjustments	INESC TEC
EXP-13	AI-enhanced control strategy for production environment	ART-ER
EXP-14	A Smart Predictive Maintenance Toolbox for drawing lines of car body element	AIN
EXP-15	Water Leakage detection	University of Maribor - UM
EXP-16	IDSS for predictive quality assurance	COMET
EXP-17	AI-enabled Platform for Zero Defect	TECNALIA



Table 2 presents data summary of AI REGIO experiments, clustered by application field.

Table 2: AI REGIO Experiments Data Summary

Data Summary for AI REGIO Experiments	
“Product Engineering and Lifecycle management” Cluster (resp. CEA)	
Purpose of the data collection/generation and its relation to the objectives of the project	The data will feed AI algorithms
Types and formats of data that will be generate/collected	<ul style="list-style-type: none"> - LPKF design rules : pdf format(Exp3) - IPCs for conception : <ul style="list-style-type: none"> o IPC2221: as pdf format(Exp3) o IPC2222 : as pdf format(Exp3) FMCEA like datasets, Excel files (Exp2)
Possible re-use any existing data and how	Yes or No (e.g. after anonymization and aggregation)
Origin of the data	S2P, Intellimech
Expected size of the data	n/a
Data utility: to whom might it be useful	n/a
“Factory Efficient and sustainable manufacturing” Cluster (Resp. EURECAT)	
Purpose of the data collection/generation and its relation to the objectives of the project	The data will be used to feed the optimization algorithms and machine learning algorithms
Types and formats of data that will be generate/collected	Process data and machine data , HDF5 (Exp17) ROS type messages (.msg files).(Exp12) ROS type services (.srv files).(Exp12)



	<p>2D Images (.jpeg /.png files) (Exp12)</p> <p>Process Data, JSON (Exp12)</p> <p>Process data, external open data (Exp10)</p> <p>Manufacturing process, CSV or XML (Exp5)</p>
Possible re-use any existing data and how	<p>Available upon request.(Exp12)</p> <p>Not Applicable (Exp17)</p> <p>No (Exp05)</p> <p>To be defined (Exp 10)</p>
Origin of the data	Tecnalía, Hohner, Inesc tec, ARMAC
Expected size of the data	<p>34 Gbytes (Exp17)</p> <p>From 5KB to 100MB(Exp12)</p> <p>3-10 MB (Exp05)</p> <p>~1GB (Exp10)</p>
Data utility: to whom might it be useful	n/a
Quality control and predictive maintenance (Resp. NISSATECH)	
Purpose of the data collection/generation and its relation to the objectives of the project	The data will feed machine learning algorithms
Types and formats of data that will be generate/collected	<p>Radar Data, format: Polarimetric, VOL.(Exp15)</p> <p>Process data, opc-ua, parquet (Exp07)</p>
Possible re-use any existing data and how	<p>Yes or No (e.g. after anonymization and aggregation)(Exp15)</p> <p>No(Exp07)</p>
Origin of the data	UM(Exp15), Arculus(Exp07)
Expected size of the data	<p>8GB(Exp15)</p> <p>~ 10 GB(Exp07)</p>
Data utility: to whom might it be useful	n/a
“Robotics and human interaction” Cluster (Res. STIIMA)	



Purpose of the data collection/generation and its relation to the objectives of the project	The data will feed AI algorithms
Types and formats of data that will be generate/collected	Production logistics data, XML (Exp01) Machine (resource) data, XML and owl (Exp11) Sensor Data, csv (Exp06)
Possible re-use any existing data and how	Yes after anonymization and aggregation (Exp01) Yes (Exp11) Yes after anonymization and aggregation (Exp06)
Origin of the data	Gualini (Exp01) TAU(Exp11) Kautenburger(Exp06)
Expected size of the data	Na(Exp01) 100MB(Exp11)
Data utility: to whom might it be useful	n/a

4.1.2 Data related to AI REGIO DIHs

AI REGIO DIHs' data collected for internal project activities will be shared within the consortium in the project data-sharing repository managed by POLIMI, in which only project consortium partners are granted access by POLIMI administrator. However, the majority of data provided by DIH aims at populating the AI REGIO portal and other community websites (as DIH4INDUSTRY for instance), hence besides being stored in the official project repository, they will be also displayed online.

In particular, as you can see from tables below, DIHs' Organization Details, Service Portfolios and Customer Pipelines are expected to be shown in the portal; while business models will remain stored in the project data-sharing repository managed by POLIMI.

As of now, we have identified four different types of data to be collected from DIHs:

- Service portfolio
- Organization details
- Customer pipeline
- Business model

And you can find all details in tables below.

Please note that, even if each DIH will provide its own information, in terms of data purpose, format, usage, size and origin, data will be provided homogeneously.



Table 3: Data Summary for DIHs Service Portfolio

Data Summary for DIHs Service Portfolio	
Purpose of the data collection/generation and its relation to the objectives of the project	<p>The Service Portfolio is the catalogue of services that the DIH currently provides (or is planning to provide in short term) to its customers. It is organized according to a 3 levels taxonomy, in order to have a common and standard structure, shared among the DIH consortium.</p> <p>The objectives of the Service Portfolio are:</p> <ul style="list-style-type: none"> • To define services provided according to a standard taxonomy, comprehensible by a larger pan-EU community of DIHs • To identify services not yet provided but to be added to the pipeline of services to be implemented, since useful to give support to customers • To share with AI REGIO consortium own capabilities and competencies, to create synergies among DIHs
Types and formats of data that will be generate/collected	<p>Currently, DIHs have been asked to fill an excel containing the taxonomy of the Service Portfolio, specifying services already provided and those to be implemented in short term.</p> <p>Moreover, the same information have been collected in an online form, in preparation to the online AI REGIO portal that will show interactively each DIH's portfolio</p>
Possible re-use any existing data and how	Data will be used in next months to populate the AI REGIO portal with DIHs information and to define the business models of service not yet provided (but to be implemented)
Origin of the data	Excel and online form manually populated
Expected size of the data	Each portfolio in excel format is about 30 KB
Data utility: to whom might it be useful	<p>The Service Portfolio is useful for:</p> <ul style="list-style-type: none"> • The DIH itself to describe services provided according to a standard taxonomy • The DIHs community to identify competences and services offered by the others (and possibly to create synergies)



Table 4: Data Summary for DIHs Organization Details

Data Summary for DIHs Organization Details	
Purpose of the data collection/generation and its relation to the objectives of the project	Organization details of each DIH are required to add the DIH into AI REGIO portal and DIH4INDUSTRY portal. In order to provide standard information, it has been prepared a template to be manually filled, with legal information and areas of competence (according to a common taxonomy)
Types and formats of data that will be generate/collected	Currently, DIHs' organization information have been collected in an online form
Possible re-use any existing data and how	Data will be used in next months to populate the AI REGIO portal with DIHs information.
Origin of the data	Online form manually filled
Expected size of the data	/
Data utility: to whom might it be useful	With the aim of creating an online portal where SMEs, DIHs, technology providers and other people will navigate searching for services and competences, to organize organization details according to a common standard is fundamental to facilitate the "search" and "browse" activities

Table 5: Data Summary for DIH Customer Pipelines

Data Summary for DIHs Customer Pipelines	
Purpose of the data collection/generation and its relation to the objectives of the project	The customer pipeline is a graphical representation of the customer journey that a specific customer profile does when contact a DIH. It consists of the list of services provided to him/her, framed in a timeline in order to understand dependencies and timing. The compilation of customer pipelines (one for each customer profile supported by the DIH) should be mainly based on success stories and experiences
Types and formats of data that will be generate/collected	Currently, DIHs have been asked to fill a power point template, defining a pipeline for each customer profile supported.



Possible re-use any existing data and how	One it will be implemented, the same information will be available also in AI REGIO portal and stored in the website database. (Each instantiated service will show as metadata the typology of the customer to whom it is addressed and the stage of the customer transformation when it is provided)
Origin of the data	Currently power point manually filled. In future data displayed online
Expected size of the data	For each DIH, the power point containing all pipelines is about 500 KB
Data utility: to whom might it be useful	At power point level, defining the customer pipelines is a useful exercise to understand if some services are missing in the portfolio to provide a complete journey to customers. At the portal level, it will be useful to show success stories to the community of DIHs and to other customers and to provide an idea of what the transformation journey will be.

Table 6: Data Summary for DIHs Business Models

Data Summary for DIHs Business Models	
Purpose of the data collection/generation and its relation to the objectives of the project	The DIHs' business models will be implemented referring to those services of the portfolio not implemented yet, in order to define specs, requirements, resources needed, timeline... Each DIH will provide a business model for each service that it is planning to implement during AI REGIO project
Types and formats of data that will be generate/collected	The activity is not started yet. No specs about business model format has been defined
Possible re-use any existing data and how	/
Origin of the data	/
Expected size of the data	The activity is not started yet. No specs about business model format has been defined
Data utility: to whom might it be useful	Business model contains information and list of activities to be followed to implement new services in an achievable way. It is the "track" to be followed.



4.2 Scientific publications Data

As outlined in section 3, AI REGIO participates in the European Commission's Open Research Data Pilot (ORD) pilot, which applies mostly to the data needed to validate the results presented in scientific publications. Open access implies unrestricted online access to research outputs such as journal articles, without access fees. The goal of the EU with this program is fostering access to and re-use of data generated by EU funded projects in order to improve and maximize public financial European resources and avoid duplication of efforts.

Participating in the ORD Pilot does not necessarily mean opening up all your research data. Rather, the ORD pilot follows the principle "*as open as possible, as closed as necessary.*"

Further procedures as regards the management of AI REGIO open access to scientific publications and research data are laid down in Article 29.2 of the Grant Agreement.

GA Article 29.2

Each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results.

In particular, it must:

(a) as soon as possible and at the latest on publication, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications;

Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications.

(b) ensure open access to the deposited publication — via the repository — at the latest:

(i) on publication, if an electronic version is available for free via the publisher, or

(ii) within six months of publication (twelve months for publications in the social sciences and humanities) in any other case.

(c) ensure open access — via the repository — to the bibliographic metadata that identify the deposited publication.

The bibliographic metadata must be in a standard format and must include all of the following:

- the terms "European Union (EU)" and "Horizon 2020";*
- the name of the action, acronym and grant number;*
- the publication date, and length of embargo period if applicable, and*
- a persistent identifier.*

4.2.1 Publishing infrastructure for Open Access

With respect to AI REGIO publications, publications can be published in open access journals best suited to the publication at hand. This option is encouraged to all project partners. This means the journal grants open access by definition. These are normally:

- Conference proceedings

- Open access journals with no Article Processing Charges (APC);
- Open access journals with "low cost" APCs

As stated in the Description of Action Open access to AI REGIO peer reviewed scientific publications will be provided with the highest standard (gold) when possible, that means a higher fee is paid and the publication is granted open access immediately.

Publications can also be published in green open access mode, that means the publication can be made available after an embargo period, which is set up by the publisher.

4.2.1.1 Publishing Process: exploring the Open Research Europe” scientific publishing service

The Open Research Europe scientific publishing platform is a free publishing infrastructure offered by the European Commission to all project publications stemming from Horizon 2020 funding projects. Partners may use this platform for rapid and transparent publishing and to facilitate open, constructive research discussion.

All articles are published using a fully transparent, author-driven model; the authors are solely responsible for the content of their article. Invited peer review takes place openly after publication, and the authors play a crucial role in ensuring that the article is peer reviewed by independent experts in a timely manner. Articles that pass peer review will be indexed in a number of bibliographic databases and repositories, following formal approval by these services.

The detailed process for Open Research Europe publications is illustrated in the graphic below:



Figure 1 - Open Research Europe Publishing Process

AI REGIO partners may decide themselves whether to use these services for their scientific publications.



4.2.1.2 Zenodo

The repository for AI REGIO open access publications will be based on the free zenodo platform (zenodo.org). Zenodo is an open access repository for scientific publications and is hosted by CERN which has existed since 1954 and currently has an experimental programme defined for the next 20+ years. CERN is a memory institution for High Energy Physics and renowned for its pioneering work in Open Access. Zenodo is funded by the European Commission via the OpenAIRE projects: FP7: OpenAIRE (246686), OpenAIREplus (283595); Horizon 2020: OpenAIRE2020 (643410), OpenAIRE-Connect (731011) and OpenAIRE-Advance (777541).

Each partner is required to create an account in order to upload respective Open Access publications, Open Access datasets or a validated public deliverable. In order to comply with Article 29 of the AI REGIO Grant Agreement each partner shall deposit at the same time the research data needed to validate the results presented in the deposited scientific publication.

AI REGIO is a “community” within zenodo where all scientific publications will be uploaded.

Publications and data uploaded on zenodo are indexed automatically in the OpenAire aggregator and will be automatically synchronized with the European Commission participant portal, hence there is no need to update information on scientific publications for reporting purposes.

In addition to zenodo, project partners might want to use other platforms to upload data or publications: Further repositories for data are available within the “re3data” registry (www.re3data.org) and OpenAire (www.openaire.eu). Project partners are encouraged to disseminate outputs also in web-based scientific social networks such as www.academia.org or www.researchgate.net

4.3 AI REGIO Deliverables Data

AI REGIO deliverables belong to two categories: Confidential and Public. Public deliverables deal with methodologies, plans of activities or descriptions of components and ICT platforms for which, in general, information content does not include datasets. On the other hand, as Confidential deliverables are inherently closed to the public, only the datasets that will be made available by the project SME manufacturers or the DIHs can be considered

5 Making AI REGIO DATA Findable Accessible Interoperable and Re-usable (FAIR)

The European Commission has highlighted the importance of making the data produced by European-funded projects **Findable, Accessible, Interoperable and Reusable (FAIR)** (<https://www.go-fair.org/fair-principles/>, s.d.)¹, with a view to ensuring its sound management, as well as boosting the dissemination of relevant information and the easy exchange of data within the EU state members. Thus, European FAIR data approach implements standards and metadata to make data discoverable, specifying data sharing procedures and which data will be open, allowing data exchange via open repositories as well as facilitating the reusability of the data.

¹ <https://www.go-fair.org/fair-principles/>



The following sections of the DMP lay out the methodology followed in the context of AI REGIO with respect to making data findable, accessible and interoperable, as well as ensuring their preservation and open access, with a view to increasing its re-use.

5.1 Making data findable

AI REGIO emphasizes the need on improving the discoverability of data produced/used during its activities. Following a metadata-driven approach to improve the searchability of data, while at the same time supporting its interpretation and re-use both for humans and computers.

To this end, project data can be identified with rich metadata relevant to its content and format and “machine-readable” to ensure automatic discovery of datasets and services. The project uses metadata that follow a globally unique and persistent identification mechanism for the development of rich and reliable metadata to promote the long-term discovery, usage and integrity of its data.

AI REGIO is expected to deposit part of generated and collected data in Zenodo, the chosen open online research data repository. Since the DMP is expected to mature during the project, the subsequent releases of the deliverable (namely some internal releases and at least the public and final report named “D1.19: Data Management Plan - Final version” due by M36) will specify the repositories where the data will be stored and go into more detail on how this data can be accessed by the wider research community.

5.2 Naming convention

In AI REGIO each data-source will be provided with a specific name that is composed by different parts/elements, containing information about the Experiment, data type or format and naming structure as follows:

ORIGIN ORG TOD FORMAT Info VERSION

The above syntax contains the following elements:

- ORIGIN: A prefix denoting the Experiment or GEN if the dataset is Experiment agnostic.
- ORG: The acronym of the related Partner.
- TOD: The type of data.
- FORMAT: The data format/extension.
- Info: Additional (abbreviated) information about the dataset. For example, the year when the dataset was published or the name of the responsible partner.
- VERSION: The version of the dataset.

5.2.1 Search keywords

The project's data will be provided with easy-to-use search keywords in order to maximize its re-use by interested stakeholders throughout project's lifecycle following the authorization procedures described in section 6. With that in mind keywords, as a subset of metadata, are used to add valuable information to the data collected/generated facilitating its discoverability and correlation to the AI REGIO project.

In this regard, the project strategy on keywords is based on the following principles:

- The who, the what, the when, the where, and the why should be covered.
- Consistency among the different keyword tags needs to be ensured.
- Relevant, understandable and explicit keywording should be followed.



5.2.2 Versioning

Versioning makes a revision of datasets uniquely identifiable, thus, enabling to keep track of the work done. More specifically, data versioning is used to define whether and how data changed over time, as well as to explicitly identify which version the creators / editors are working with. In addition, effective data versioning makes it easier to understand whether an updated version of a dataset is available and which changes are made between the different versions, allowing comparisons and avoiding confusion. In this context, a clear version number indicator is used in the naming convention of every data file produced during AI REGIO in order to facilitate the identification of different versions.

5.3 Making data openly accessible

In order to maximize the impact of AI REGIO project, research data will be shared within and beyond the Consortium. Selected data and results will be shared with the scientific community and other stakeholders through publications in scientific journals and presentations at conferences, as well as through open access data repositories. All data are made available for verification and re-use.

In AI REGIO data will be findable and accessible by providing a common repository for storing the data and offering a simple programming interface for accessing it.

Public deliverables listed in the GA will be made publicly available via the AI REGIO website. All personal and sensitive information will be removed from these datasets/ reports before they are made public.

5.4 Making data interoperable

It is essential that published datasets are unequivocally interpretable by third persons without any link to the project. Therefore, each dataset needs to be accompanied with a description of the methodology, sources, definitions and scope of the data contained in it. Whenever possible, datasets should be structured in such a way that it can, in full or in part, be combined with another dataset, from the project or any other data source.

AI REGIO will adopt in its data management methodology the use of metadata vocabularies, standards, and methods that will increase the interoperability of the data collected/generated through its activities.

More specifically, standard vocabularies will be used for all data types present in the project. In case there is an uncommon vocabulary, a clear mapping will be provided in order to facilitate its use. Thus, the project's data will be interoperable and easy for sharing among researchers, institutions and organizations.

Partners will observe the OpenAIRE guidelines for online interoperability, available [here](#) and they will also ensure that AI REGIO data observes FAIR data principles available [here](#).

As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP.

5.5 Making data reusable

AI REGIO is expected to produce a substantial volume of data and knowledge that will be presented to the scientific community, to the Digital Innovation Hubs community, industry, policy-makers and society at large through a carefully designed portfolio of dissemination actions. Each dataset will have an individual license. Access to project data will be provided to the whole AI REGIO Consortium and exclusively for the project objectives. Datasets produced as a result of the project work will be



shared within the Consortium and will only be allowed for external sharing with a consensual Consortium approval of the relevant stakeholders, by accepting the terms and conditions of use, as appropriate. The license for the access, sharing and re-use of AI REGIO material and output datasets will be defined by the Consortium on a case-by-case basis. As the project progresses and data is identified and collected, further information on increasing data re-use will be outlined in subsequent versions of the DMP. In specific, information on how data will be licenced to permit the widest reuse possible, when the data will be made available for re-use, whether the data produced and/or used in the project is useable by third parties, a description of data quality assurance processes and specifications of length of time for which the data will remain re-usable will be provided.

In AI REGIO data will be made interoperable and reusable by:

- Designing semantic models and domain ontologies for manufacturing and human centred AI that are shared across all the pilots
- Reusing of existing relevant standards such as RAMI 4.0, AAS, OPC-UA, Automation ML, B2MML (further details on the relevant standards will be reporter within WP5 scope).

6 Data Security

AI REGIO will securely handle any collected/generated data throughout its entire lifecycle as it is essential to safeguard this data against accidental loss and/or unauthorised manipulation. Particularly, in case of personal data collection/generation it is crucial that this data can only be accessible by those authorised to do so. With that in mind, the project's backup and data recovery strategy aims at ensuring that no data loss will occur during the course and after the completion of AI REGIO, either from human error or hardware failure, as well as inhibit any unauthorised access.

All project partners are responsible for processing data within their private servers and will ensure that this data is protected, and any necessary data security controls have been implemented, to minimise the risk of information leak and destruction. This case refers to the data that will be closed and therefore will not be shared and/or re-used within the framework of the project.

The security principles are listed below:

- **Authentication:** All the users wanting to get access to the AI REGIO datasets should be authenticated. Also, proper means are used to authenticate the servers. An authentication system could be used to handles the authentication of the users during the course of the project.
- **Authorization:** The access to AI REGIO datasets must be only available to the authenticated and authorized users. These categories and the rights of those users are defined and enforced. The appropriate access control policies and mechanisms (including physical access control) shall be identified for each Experiments.
- **Accounting:** In AI REGIO any access and modification to a resource by any user is securely traced/logged in order to prevent users from denying that data files were accessed, altered or deleted, auditing. Other accounting mechanisms shall be implemented.
- **Confidentiality:** The data produced in AI REGIO should be encrypted during transmission and potentially even in storage.
- **Communication Security:** Access to AI REGIO portal and repository should be done through encrypted communication channels such as HTTPS and IPsec.
- **Data Integrity:** The data collected during AI REGIO should be protected from malicious and accidental modifications by any users during their transmission or their storage.



All the AI REGIO tools and platforms (e.g. the AI REGIO AI4Manufacturing Toolkit, the Data4AI Platform, the DIHIWARE, ...), as well as any other integrated platform, will ensure the respect of these security principles through:

1. **User login:** Account information will be protected by a password. The password for the account will exist on the platform only in a coded form and AI REGIO will not store the non-encrypted version. Furthermore, envisaged deployments of digital solutions powered by AI REGIO will use SSL certificates;
2. **OAuth2** – AI REGIO will foster the adoption of standard implementation of RFC 6749 – token-based authorization layer for control of client access to resources;
3. **API keys**– The public API of the back-end services and databases of AI REGIO will be accessible through an API key, which can be created and managed by registered users;
4. **Encrypted cookies** – Front-end cookies containing session information are exchanged between the web UI and the backend. This cookie stores a session identifier and encrypted session data when users are logged in to the AI REGIO web applications.

Security will be considered additionally for the purposes of data exchange between partners and sharing before the final data integration/publication, and further specifications and details will be provided as part of task T5.2. The particular security measures will be taken on a case-by-case basis based on the medium for data exchange and the precise needs of each data provide. They will include the following:

1. Setting up security policies on cloud service providers;
2. Setting up secure FTP server for file transfer of any files over the Internet;
3. Setting up secret SSH keys for accessing servers/clusters of servers with running databases that host any shared dataset.

In case of a catastrophic event that implies the partial or complete deletion of the datasets, the data from the most recent back up will be restored.

Identification and authentication access/usage controls play an important role in the context of the project, as they help partners to protect the data produced/used during AI REGIO and especially personal data. To this end, each project partner is responsible for and committed to ensuring the application of appropriate access controls to the data they are processing within their private servers of their organisation. Moreover, in order to safeguard the privacy of the users of the project's Web applications and AI REGIO tools and platforms, dedicated privacy policies will define the way in which these online spaces collect, process and use personal data, the security procedures followed, the users' rights as well as the cookies policy employed.

7 Storage and Preservation of Data

Given the possibility for the AI REGIO project to manage high volumes of data, a plan for storage and backup of data must be realized, in order to foresee situations of exponential growth of the data volume or data loss.

7.1 Storage

This section provides initial considerations for the AI REGIO data storage in alignment with the FAIR data principles and Open Research Data (ORD) pilot. Since the DMP is expected to mature during the project, the subsequent releases of this deliverable will better specify further also the storage policies.

Data in the AI REGIO will be exchanged and made available through a two-tier storage policy. The policy will consist of:



- Tier 1: a shared data space for exchanging raw input data between Consortium partners.
- Tier 2: a structured data storage with integrated datasets, interoperable with the AI REGIO AI4Manufacturing Toolkit and Data4AI Platform, will be used to produce the integrated data according to a shared template.

Tier 1 will be implemented using a private file or data sharing solution. It will use cloud hosting infrastructure services to enable easy access over the web. Data will be stored using data hosting service and secure data sharing protocols to ensure that data are not compromised.

Tier 2 will be implemented based on specific solutions and platforms provided by the AI REGIO Consortium as part of the AI4Manufacturing Toolkit and Data4AI Platform software stack, where the shared data models will be published and the output data will be imported in dedicated database management system(s) and registered in a virtual catalogue, taking into account the user access restrictions for each dataset.

7.1.1 Backup and Recovery

Backup and recovery strategy is specified and planned for each dataset in order to prevent data loss risks. Confidentiality must be strictly maintained, and anonymization is applied where required. In addition, partners are not allowed to reveal sensitive data about participants to test cases, pilots, trials etc. These same principles are also being taken into consideration in the dissemination of data. Input datasets have already backup and recovery in place (when needed) under the responsibility of the data provider(s) and are directly managed by them; therefore, no backup and/or recovery mechanism for input datasets falls within the scope of the AI REGIO ecosystem.

The concrete data backup and recovery mechanisms to be adopted at AI REGIO ecosystem level will be discussed in the future versions of the Data Management Plan as they evolve throughout the project, or in other deliverables dealing with technical aspects (such as the detailed design of the platform or the business cases implementation plans).

7.1.2 Data Archiving

The data used and produced during the project development will be updated each time they change in project lifetime. For each dataset update, a reference documentation will also be produced. This supporting documentation will report the changes of the dataset respect to previous version.

AI REGIO datasets used in the demonstrator will be maintained for at least four years after project termination. Sensitive data preservation will follow the guidelines that AI REGIO Consortium will provide during the project development.

7.2 Preservation

Data preservation refers to data upkeep and maintenance to ensure that the integrity of the data is upheld in the future. This includes properly maintaining the data repository and data backups to ensure the long-term value of the data.

The data used and produced during the project development are updated each time they change in project lifetime. The newly updated datasets are identified by the number of the previous dataset version plus one, according the naming convention reported above. For each dataset update, a reference supporting documentation is also produced. The document reports changes of the dataset respect to the previous version. Also, the reference document is identified by the number of the previous document version plus one, according the naming convention reported.

As a project funded by the European Commission, AI REGIO datasets will be maintained for at least four years after project termination. In this time, the Consortium will ensure that they remain accessible and usable, and destruction of research data will not take place, unless a participant



requests it. For cases in which the Consortium will not be able to keep data available for the established time, specific archiving policies for those data will be provided and well documented by the Consortium.

Project data can be deleted only after that related archiving deadlines are reached. Destruction of research data before established deadlines will not take place, unless a participant formally requests it and the Consortium approves the deletion. Additionally, deletion of personal data will be carried out upon triggering of the situations provided by Art. 17 of the GDPR (Right to erasure) (Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR)).² In this case, upon request from the data subject (as defined in the above GDPR article), dataset owner / managing partners (acting as data 'controller') will abide to the GDPR provisions and follow suit.

8 Data Quality Assurance

Data quality assurance is the process of data profiling to discover inconsistencies and other anomalies in the data, as well as performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality (Redman, 2013).

Data quality refers to the state of qualitative or quantitative pieces of information. Data is generally considered high quality if it is "fit for [its] intended uses in operations, decision making and planning" (Liability and Leverage - A Case for Data Quality, August 2006. Archived from the original on 2011-01-27. Retrieved 2010-06-25.). Moreover, data is deemed of high quality if it correctly represents the real-world construct to which it refers (https://en.wikipedia.org/wiki/Data_quality).

People's views on data quality can often be in disagreement, even when discussing the same set of data used for the same purpose. When this is the case, data governance is used to form agreed upon definitions and standards for data quality. In such cases, data cleansing, including standardization, may be required in order to ensure data quality.

Our approach for data quality assurance will be based on the work in WP4 and WP5, Task 4.2 and Task 5.3. The goal of the approach is to enable an efficient (time, resources) and affordable (experts, costs) approach for ensuring data quality.

Regarding AI applications, the most important criteria is the suitability of the data (content) for resolving AI-problems. We will create corresponding methods for a) measuring initial data quality (data source) and b) determining the desired quality of the "clean" data, serving as inputs in AI-based learning processes

9 Ethical Aspects and Legal Compliance

All the aspects related to legal compliance and ethical issues have been deeply described in D1.11 "Ethics Assessment and Data Management Strategy", which also draws the roadmap for the implementation of legal and ethics related activities within the project. Such document outlines AI REGIO Ethical Strategy, the main ethical procedures and responsibilities for human participation in the experiments and data protection policy. It also identifies the oversight responsibilities and provides the basis for the development of the Ethics and Data Protection Impact Assessment methodology to be used within Work-Package 6, as well as contains the findings on how the indicators for prioritizing human well-being and flourishing in AI REGIO will be defined. Reference should be made to such deliverable for further details.

² See <https://gdpr-info.eu/art-17-gdpr/>



10 Allocation of resources

The project consortium includes resources to make data FAIR- compatible, therefore no extra costs are expected. In order to properly tackle with FAIR paradigm, each of the experiments appointed a Pilot Data Manager. They are as shown in the table below.

Table 7: AI REGIO Pilot Data Managers

Role	Organization Name	Representative name
Pilot Data Manager EXP01	GUALINI	Germana Gianquinto
Pilot Data Manager EXP02	INTELLIMECH	Fabio Floreani
Pilot Data Manager EXP03	S2P	Mael MOGUEDET
Pilot Data Manager EXP04	SWARM	Ramzi ABBES
Pilot Data Manager EXP05	HOHNER	Edgar Rubion Soler
Pilot Data Manager EXP06	KAUTENBURGER	Kevin Schommer
Pilot Data Manager EXP07	ARCULUS	Carlo Fitz
Pilot Data Manager EXP08	BPIND	John Blankendaal
Pilot Data Manager EXP09	SKU	Francisco Souza
Pilot Data Manager EXP10	SKU	Francisco Souza
Pilot Data Manager EXP11	TAU	Eeva Järvenpää
Pilot Data Manager EXP12	INESC TEC	Rafael Arrais
Pilot Data Manager EXP13	ART-ER	Domenico Guida
Pilot Data Manager EXP14	AIN	Jorge Alonso Marquez
Pilot Data Manager EXP15	MARIBORSKI VODOVOD	Bojan Erker
Pilot Data Manager EXP16	COMET	Riccardo Zanelli
Pilot Data Manager EXP17	TECNALIA	Cabero Lopez, Jose Maria

11 AI REGIO Datasets

This section provides an explanation of the different types of datasets to be produced or collected in AI REGIO, which have been identified at this stage of the project. As the nature and extent of these datasets can evolve during the project, more detailed descriptions will be provided in the next version of the DMP towards the end of the project.

The next release of the document will also contain the datasets list.



11.1 Datasets categories

Initial categorization of the datasets is based on the existing clustering of experiments. Therefore, we propose following categories of datasets:

- Product Engineering and Lifecycle Mgmt. Datasets
- Factory efficient and sustainable manufacturing Datasets
- Quality control and predictive maintenance Datasets
- Robotics and Human Interaction Datasets

11.2 Update of AI REGIO Dataset templates

This section provides an explanation of the different types of datasets to be produced or collected in AI REGIO, which have been identified at this stage of the project. As the nature and extent of these datasets can evolve during the project, more detailed descriptions will be provided in the next version of the DMP towards the end of the project.

The next release of the document will also contain the datasets list.

11.2.1 General template

Table 8: General Template

General Dataset Template	Description	Mandatory / Optional	Value
Title	Name of the dataset	Mandatory	
Creator	Entity primarily responsible for making the dataset	Mandatory	
Description	Overview of the dataset, CS associated.	Mandatory	
Type	Data type	Mandatory	
Purpose	Purpose of the dataset: to optimize, to monitor, ...	Mandatory	
Main Use	When is its main use in the project WPX Task X.X	Mandatory	
Personal data	Does it contain personal information (and what types)	Mandatory	
Initial Availability Date	Initial date of available data	Optional	
End Date	End date of data collection	Optional	
Availability	Availability of the dataset is public or private	Optional	
Source URL	URL Location of the dataset	Optional	
Access restrictions	Restrictions users and passwords	Optional	
Recovery	Data recovery methods and secure storage	Optional	
Privacy management	Procedures of privacy management	Optional	



Language	Language of the dataset	Optional	
Relation	Relation with other datasets	Optional	
Owner	Name of the Partner	Mandatory	
Contact Person	Contact Information	Optional	
Relation to the project	Role of the person	Optional	
License	Public license if the "Access Level" is public or internal terms and conditions	Mandatory	
Version	Version of the dataset		

11.2.2 Technical template

Table 9: Technical Template

Technical Dataset Template	Description	Mandatory / Optional	Value
Structure	Structure of the dataset (column names and column types)	Mandatory	
Type and Format	File format of the dataset	Mandatory	
Data Size	Size of the dataset	Optional	
Important features	Most important parameters	Optional	
Data Rate	Data collection sampling rate	Optional	
Communication protocol	Protocol of devices and sensors	Optional	
Edge or embedded	Edge or embedded elaboration	Optional	
Tools to access	A related hardware, tools, allow to the access	Optional	
Name of the system	Name of the system that is the main source of the dataset	Optional	
Visualisation	Ability to view the data using HMI, SCADA, ...	Optional	
Technology	Technology adopted to expose the dataset	Optional	
Source	Which tools were used to collect the data	Optional	
Destination	What is the destination of the data	Optional	
Archiving	Where data are exploited, stored (SCADA, MES, ERP) cloud database, embedded computational unit, control software...	Mandatory	



12 References

<https://www.go-fair.org/fair-principles/>. (s.d.).

Liability and Leverage - A Case for Data Quality. (August 2006. Archived from the original on 2011-01-27. Retrieved 2010-06-25.). *Information Management*.

Müller, V. C. (2020). "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopaedia of Philosophy*. Edward N. Zalta.

Redman, T. C. (2013). Data Driven: Profiting from Your Most Important Business Asset. *Harvard Business Press*. ISBN 978-1-4221-6364-1.

Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR).