Common European Data Spaces and Robust AI for Transparent Public Governance

CEDAR

Project acronym: CEDAR

Project full title: Common European Data Spaces and Robust AI for Transparent Public Governance

Call identifier: HORIZON-CL4-2023-DATA-01

Type of action: HORIZON-RIA

Start date: 01/01/2024

End date: 31/12/2026

Grant agreement no: 101135577

D7.2 Data Management Plan

Document description: *List of datasets used and generated in the project, associated data management aspect.*

Work package: *WP7*

Author(s): *Irem Goymen, Dr. Sophia Karagiorgou*

Editor(s): *UPM, ENG, KUL, ALBV, DBC, INS, LC, ANCE, SBC, SNEP, MDP, MNZ, VICOM, ART, UC, CNT*

Leading partner: *UBI*

Participating partner: *ALBV, CEA*

Version: *1.0*

Status: *Submitted*

Deliverable type: *DMP – Data Management Plan*

Dissemination level: *PU - Public*

Official submission date: *01/07/2024*

Actual submission date: *28/06/2024*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101135577

Disclaimer

This document has been produced in the context of the CEDAR Project. This project is part of the European Union's Horizon Europe research and innovation programme and is, as such, funded by the European Commission. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. All information in this document is provided 'as is' and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability with respect to this document, which is merely representing the authors' view.

The CEDAR consortium consists of the following partners:

No.	Partner Organisation Name	Partner Organisation Short Name	Country
1	Centre for Research and Technology Hellas	CERTH	Greece
2	Commissariat al Energie Atomique et aux Energies Alternatives	CEA	France
3	CENTAI Institute S.p.A.	CNT	Italy
4	Fundacion Centro de Technologias de Interaccion Visual y Comunicaciones VICOMTECH	VICOM	Spain
5	TREBE Language Technologies S.L.	TRE	Spain
6	Brandenburgisches Institut für Gesellschaft und Sicherheit GmbH	BIGS	Germany
7	Christian-Albrechts University Kiel	KIEL	Germany
8	INSIEL Informatica per il Sistema degli Enti Locali S.p.A.	INS	Italy
9	SNEP d.o.o	SNEP	Slovenia
10	YouControl LTD	YC	Ukraine
11	Artellence	ART	Ukraine
12	Institute for Corporative Security Studies, Ljubljana	ICS	Slovenia
13	Engineering – Ingegneria Informatica S.p.A.	ENG	Italy
14	Universidad Politécnica de Madrid	UPM	Spain
15	Ubitech LTD	UBI	Cyprus
16	Netcompany-Intrasoft S.A.	NCI	Luxembourg
17	Regione Autonoma Friuli Venezia Giulia	FVG	Italy
18	ANCEFVG – Associazione Nazonale Construttori Edili FVG	ANCE	Italy
19	Ministry of Interior of the Republic of Slovenia / Slovenian Police	MNZ	Slovenia
20	Ministry of Health / Office for Control, Quality, and Investments in Healthcare of the Republic of Slovenia	MZ	Slovenia
21	Ministry of Digital Transformation of the Republic of Slovenia	MDP	Slovenia
22	Celje General Hospital	SBC	Slovenia
23	State Agency for Reconstruction and Development of Infrastructure of Ukraine	ARU	Ukraine
24	Transparency International Deutschland e.V.	TI-D	Germany
25	Katholieke Universiteit Leuven	KUL	Belgium
26	Arthur's Legal B.V.	ALBV	Netherlands
27	DBC Diadikasia	DBC	Greece
28	The Lisbon Council for Economic Competitiveness and Social Renewal asbl	LC	Belgium
29	SK Security LLC	SKS	Ukraine
30	Fund for Research of War, Conflicts, Support of Society and Security Development – Safe Ukraine 2030	SU	Ukraine
31	ARPA Agenzia Regionale per la Protezione dell' Ambiente del Friuli Venezia Giulia	ARPA	Italy

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.1	30.04.2024	ToC regarding the HE DMP Template	UBI, Dr. Sofia Karagiorgou
0.2	06.05.2024	First Draft	UBI, Irem Goymen
0.3	25.05.2024	Consolidated Draft; sent for internal review	UBI, Dr. Sofia Karagiorgou
0.4	10.06.2024	Internal review 1	ART, Dimitra Stefanatou
0.5	11.06.2024	Internal review 2	CEA, Mounir Kellil
1.0	24.06.2024	Final review before submission	CERTH, Thodoris Semertzidis and Maria Anastasia Minopoulou
1.1	26.06.2024	Final version	UBI, Irem Goymen

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise.

Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1 Introduction 7
1.1 Purpose of the Document 7
1.2 Relation with Other Tasks and Deliverables 7
1.3 Deliverable Structure 8
2 CEDAR Data Management Overview 9
2.1 CEDAR Datasets 9
2.2 Data Origin 10
2.2.1 Pilot 1: Transparent Management of National RRP Funds in Italy [led by INS in T5.1] 11
2.2.2 Pilot 2: Transparent Management of Slovenian Public Healthcare Funds [led by ICS in T5.2] 11
2.2.3 Pilot 3: Transparent Management of Foreign Aid for Rebuilding Ukraine [led by YC in T5.3] 12
2.3 Types and Formats of Artefacts Generated / Collected 13
2.4 Project Artefacts and Access Rights 13
2.5 Possible Reuse of the Data 16
2.6 Expected Size of the Data (if known) 16
2.7 Data Utility 16
3 FAIR Data 18
3.1 Making Data Findable 18
3.1.1 Data Identification Mechanisms 19
3.1.2 Naming Conventions Used and Versioning 19
3.1.3 Standards for metadata creation (if any) 20
3.2 Making Data Accessible 20
3.3 Making Data Interoperable 21
3.4 Making Data Reuseable 21
4 Other Research Outputs 23
5 Allocation of Resources 24
6 Data Security 25
7 Ethics 26
8 Conclusions 28
9 References 29
10 Appendix A – Dataset List 30
10.1 Pilot 1 – Transparent Management of National RRP Funds in Italy 30
10.2 Pilot 2 – Transparent Management of Slovenian Public Healthcare Funds 31
10.3 Pilot 3 - Transparent Management of Foreign Aid for Rebuilding Ukraine 33
10.4 Other Datasets 35

List of Figures

Figure 1. CEDAR Artefacts Distribution..... 14
Figure 2. Open Access to Scientific Publication and Research Data [22]. 20
Figure 3. DMP Register. 36

List of Tables

Table 1. CEDAR Dataset Template. 10
Table 2. Artefacts Overview. 13
Table 3. CEDAR Artefacts Distribution. 14
Table 4. Partners' research item provision to CEDAR. 14
Table 5. Partners' dataset provision to CEDAR..... 15
Table 6.Metadata template for CEDAR datasets. 21

List of Terms and Abbreviations

AI	Artificial Intelligence
ALLEA	All European Academies
API	Application Programming Interface
ARU	Agency for Reconstruction and Development of Infrastructure
CA	Consortium Agreement
CCL	Creative Commons Licenses
CEDS	Common European Data Spaces
CUC	Central Purchasing Office
CV	Computer Vision
DMP	Data Management Plan
DOI	Digital Object Identifiers
DX.X	Deliverable X.X
EO	Earth Observation
EU	European Union
FAIR	Findable, Interoperable, and Reusable
FVG	Friuli Venezia Giulia
GA	Grant Agreement
GDPR	General Data Protection Regulation
HEU	Horizon Europe
IEEE	Institute of Electrical and Electronics Engineers
ML	Machine Learning
MX	Month X
NLP	Natural Language Processing
ORE	Open Research Europe
PEPs	Politically Exposed Persons
TX.X	Task X.X
WP	Work Package

Executive Summary

The Data Management Plan (DMP) of the CEDAR Project presents a detailed data management methodology for data recording, along with the measures and the policies which are applicable and aligned with the Horizon Europe (HEU) open and private data requirements. It also reports the activities that will be followed by the CEDAR Consortium to making data Findable, Interoperable and Reusable (FAIR) to the widest possible public.

The DMP details how data is collected, stored, and shared, specifying serialization and interoperability standards. It addresses data security and ethics, including compliance with the GDPR. The DMP Register acts as a living document, tracking data specifications, research, and software artifacts, ensuring alignment with open data principles.

1 Introduction

This deliverable introduces the CEDAR Data Management Plan (DMP), as collectively defined by all the Consortium Partners and registered by M6 of the project. In accordance with the European Commission's Guidelines[10] for Horizon Europe (HEU) Programme, a DMP needs to be submitted within the first six (6) months of the project. The present report constitutes the DMP of the CEDAR Project reflecting the technical progress at the moment of drafting the present document. Further updates in this respect will be provided in the course of the project's duration, on the basis of the actual developments of the technical work. For this reason, we have initiated and maintained a DMP Register, acting as a living document in the project's centralized repository. The DMP Register will be periodically updated with the latest status of collected, available and generated data, as the project progresses. Official reporting on the CEDAR's data evolution will be also provided in D7.4 and D7.6 Data Management Plan at M18 and M36, respectively.

In particular, in the direction of the principles to make research data either collected and/or generated throughout and after the project Findable, Accessible, Interoperable and Reusable (FAIR), the deliverable D7.2 details the CEDAR Consortium methodology to realize this objective. To this end, the deliverable outlines -among other- how the research data collected and/or generated will be handled during and after the CEDAR Project, describes which standards and methodologies for data identification, collection, recording and generation will be followed, and whether and how data will be shared. Note that the document is largely based on the related template provided by the European Commission.

1.1 Purpose of the Document

The Data Management Plan describes how the data will be collected, classified, stored and made FAIR within the CEDAR Project. It specifies the type of data that will be generated or collected during the project, the serialization and interoperability standards to be used, how the research data will be preserved and what parts of the datasets will be shared for verification or reuse.

The plan is expected to be updated and adjusted regularly, in line with the progress of the project. The CEDAR DMP addresses the following aspects:

- The overview of the Data Management approach in CEDAR;
- The methodology adopted to make data FAIR;
- Other research outputs that are closely monitored in DMP;
- Allocation of participating organisations and resources;
- Data security; and
- Data ethics.

1.2 Relation with Other Tasks and Deliverables

T7.2 Scientific, Technical, and Data Management with the collaboration of the Pilot and Technical Partners is responsible for the identification of datasets and the methodology to be followed during and after the project to maximise synthetic data generation, access to and re-use of data model available for the CEDAR Data Space in line with data protection and open access policies for selected datasets. This is a continuous activity with the CEDAR Project lifecycle and reporting is being performed through the living document of the DMP Registry. D7.2, will also cover the datasets that will be collected in WP1-WP5 fully adhering to the FAIR principles. Also, it reports the efforts and activities performed by the Consortium to define the baseline, the methodology and the recording frequency towards Data Management, while updates about the data evolution will be incorporated in official reporting on the CEDAR's data evolution in D7.4 and D7.6 Data Management Plan at M18 and M36, respectively.

The DMP will also outline how, where and by whom data will be stored and secured and who will have access to it. It will be integrated with T6.1 for IPR management and to opt for ethical and legal issues governed. This way, the

DMP complies with CEDAR's mission of advancing open science and data sharing, as it covers the whole data management life cycle, guaranteeing proper documentation, data curation, and data archival for future projects.

1.3 Deliverable Structure

The structure of the rest of this document is, as follows:

Chapter 2 outlines the data that have been currently identified and collected. This data will serve as the basis to initiate the technical activities of the project. Also, it includes the type and format, the purpose, the size, and finally the origin of this data. When existing data is re-used in the project, it is also stated, as well as the purpose of re-using it. In addition, the potential of the data to be used outside of the project is explained.

Chapter 3 specifies the measures to ensure the data's Findability, Accessibility, Interoperability and Reusability.

Chapter 4 details the other research artefacts and outputs that have been identified/reused in the project (e.g. software, models, studies, new materials, etc.).

Chapter 5 includes a description of the resources such as costs associated with compliance with the FAIR principles and points out who will be responsible for data management, generation, and maintenance.

Chapter 6 presents the currently identified mechanisms to ensure data security, including its storage and recovery.

Chapter 7 presents any ethical or legal issue that may have an impact on data sharing. Additionally, when the research uses personal data, aspects such as informed consent or long-term preservation is also considered.

Last, Chapter 8 concludes the deliverable.

2 CEDAR Data Management Overview

This section describes the data that will be used in the project, as identified at Month 6 (M6). To keep track and preserve data collected, generated, or enriched in the project lifecycle, we maintain and continuously monitor a DMP Register by logging data specifications, research, and software artefacts. Regarding the CEDAR's data, we keep track of the type and format, the purpose (i.e., under which work package and task, partners involved, description, usage, status), the size, and finally the origin of the data. Existing data that will be reused in the project will be described, along with a clarification of the purpose of reusing it. In addition, the potential of the data to be used outside of the project will be also explained. Besides, the DMP Register and the present deliverable both address the necessity to keep track of the purpose of the data generation or reuse, the relation with the objectives, work packages and tasks of the project. In the same direction, metadata regarding the data size, types and formats, partners involved, status, usage in CEDAR and usage beyond CEDAR are being collected. Aligned with the Open Data principles and making CEDAR's data FAIR, we monitor if the dataset is open, where it is stored, who is the owner, how will be made it open, and if it is expected to increase data utility or reuse within or outside the project and to whom. Last, in alignment with the General Data Protection Regulation (GDPR) [11], personal data protection and ethics related aspects, more broadly, are addressed including those pertaining to privacy-by-design[1] and security-by-design[2] [12] to service safeguarding their disclosure to third parties and outside the CEDAR Consortium.

2.1 CEDAR Datasets

This section presents the template of the Data Management Plan that has been set up for the collection and description of the datasets in use or foreseen in CEDAR during the reporting period. The template has been slightly adapted to the project's needs as per the actual knowledge (M6) from the Horizon Europe Template [2] and maybe will be further enriched in the future to adhere at best to project's needs. The template is presented in Appendix B – DMP Register and has been shared and explained to all partners through dedicated calls. Then, each partner autonomously listed and detailed its own datasets, got feedback from the CEDAR's Project Manager (i.e., UBI, in this case) and finalised its input in an aligned manner.

The template shared within the CEDAR Consortium includes datasets overview, details, if it is open or reused and if there are any ethics constraints. The master document is the DMP Register, has been stored in CEDAR's Repository and is a living document. The DMP register will be the reference to collect all datasets that will be used in CEDAR during the project lifetime. Any time a new dataset emerges or there are updates on an already listed one, the respective partner is responsible of enriching or updating the document.

The full list of datasets provided by the partners is reported in Appendix A through metadata specifications and without disclosing business or sensitive information; it is important to remark that being this report delivered at M6, there are still many datasets to be collected, and, in general, many details to be clarified in the next versions of this deliverable, namely in D7.4 and D7.6 Data Management Plan at M18 and M36, respectively. Table 1 presents the structure of the template that has been populated by the CEDAR Consortium for the DMP Register.

[1] *The term "Privacy by Design" means nothing more than "data protection through technology design." Behind this is the thought that data protection in data processing procedures is best adhered to when it is already integrated in the technology when created.*

[2] *Security by Design is a methodology to strengthen the cybersecurity of the organization by automating its data security controls and developing a robust IT infrastructure.*

Table 1. CEDAR Dataset Template.

Overview	Dataset ID	
	Dataset Title	
	Work Package	
	Task Deliverable	
	Partner(s)	
	Data Type	
	Data Format	
	Personal data	
Details	Description	
	Expected Size of the Data	
	Status	
	Use in CEDAR	
	Use beyond CEDAR	
	Technical and organizational measure	
Open or Reuse Data	Will you reuse any existing data? If YES, how?	
	Methodologies for data collection / generation	
	Data Storage Location	
	Metadata and Standards	
	For whom might the dataset be useful?	
	Data access, sharing and licensing	
	Security	

2.2 Data Origin

Due to CEDAR's inherent nature, data utilization is crucial in terms of ensuring the project's success. Additionally, data is expected to be generated and collected by integrating CEDAR to the various use-cases. There are three use cases in CEDAR Project, and in each of these use cases data has a different source. To avoid repetition of content, we briefly present in this document only the identified datasets, and their characteristics relating to the scope of

each use case, as they have been defined by M6. A detailed report about the pilots will be reported in D1.1 Use Cases, similarly, due at M6.

2.2.1 Pilot 1: Transparent Management of National RRP Funds in Italy [led by INS in T5.1]

In 2014, the Friuli Venezia Giulia (FVG) region established a Central Purchasing Office (CUC) with the mandate to oversee public tenders. The primary objectives of the CUC encompass the optimization of public expenditure, standardization of procurement procedures, and the preservation of legality and transparency within the public procurement arena.

To realise this pilot, several public and private data in Italian and English have been identified to be used in the project, as follows:

- Structured and unstructured data from past tenders and bidders (tabular, textual) from FVG and INS.
- Environmental sensor data from the FVG region – raw data and data from complex simulation models developed by ARPA (on air pollution, status of internal and costal water bodies, sewer systems, noise pollution, etc.).
- Satellite imagery data (Sentinel / Copernicus low-resolution and commercial high-resolution).
- Statistics and aggregate data on builder sector from ANCE.
- Data collected on site (construction site) referring to workers, tools, environmental issues, and pollutant emissions (e.g., water, air, noise) by FVG.
- Data from several public registers (e.g., National Institute for Social Security, demographic registers, health registers (occupational health), public vehicle registers).
- Data from social media (text, images, videos).

The pilot intends to enhance the eAppaltiFVG platform by integrating (1) a refined data repository comprising pertinent data from diverse domains (e.g., tenders, demographics, health, social insurance, environmental factors, vehicle/machinery registries, material specifications), and (2) a suite of AI-driven instruments facilitating meticulous oversight across all procurement stages. These tools will enable both digital and physical inspection, encouraging preventive anomaly detection and proactive response, including:

- Analysis of tenders, bids, and bidders (NLP and data mining) for monitoring procurement.
- Analysis of environmental data (CV and EO, sensor data mining) for verifying pollution reports.
- Analysis of workers' data (NLP and statistical analysis) for monitoring execution of projects.
- Analysis of social media (multi-media analysis) for capturing hints and signs of on corrupt activities.
- Analysis of cryptocurrency flows (graph-based approaches) for detecting fraud in crypto transactions.

2.2.2 Pilot 2: Transparent Management of Slovenian Public Healthcare Funds [led by ICS in T5.2]

In the realm of public procurement within Slovenia, tendering processes for low-value contracts present notable challenges. These contracts, encompassing goods and services valued below 40,000 EUR, and construction projects below 80,000 EUR, exhibit a distinct lack of stringent regulation, thereby fostering non-uniform practices. Particularly within the Slovenian healthcare domain, individual hospitals undertake the management of their respective low-value tenders, characterized by peculiar methodologies, strategies, and mostly antiquated technological infrastructures.

To realise this pilot, the pilot partners involved will utilise and enrich the following Common European Data Spaces (CEDS): Health (procurement processes in healthcare), Industrial (cost of goods), Finance (procurement transactions, financial transactions of high-risk legal entities), and Public Administration (procurement data). In this pilot, several public and private data in Slovenian and English are involved to the project, as follows:

- Structured and unstructured data from past tenders and bidders (tabular, textual) from SBC.

- Data from several public registers (e.g., ERAR – a public source of data on public procurement maintained by the Slovenian anti-corruption agency, the eNarocanje.si national portal for public procurement management).
- Data on criminal activities of legal entities and past instances of corruption from MNZ.
- Data from web (e.g., the ceneje.si portal describing up-to-date market prices of goods, data from procurement applications maintained by other Slovenian hospitals).
- Data from social media (text, images, videos).

The pilot aims to digitise the current archive of past tenders and bids that comprises documents in different formats (PDF, Word, Excel) in different locations, transform them into rich metadata, integrate them with external sources (e.g., open data, public data, CEDS), and thereby enable their analysis to identify patterns that may indicate fraudulent activities. CEDAR aims to digitalize the low-value tender procurement procedures within Slovenia's healthcare domain, thereby supporting the transparency in the management of public finances. This initiative facilitates real-time oversight of procurement activities, enabling the timely identification of potentially fraudulent or corrupt behaviour. It includes preventive detection mechanisms such as reviewing tender descriptions and post-tender analysis, comparing bids in current and past tenders as well as market standards to reduce bias in favour of certain bidders. By leveraging the latest data technologies and analytical algorithms, we strive to effectively uncover patterns and irregularities in the purchasing process, including:

- Analysis of tenders, bids, and bidders (CV, NLP, data mining) for monitoring procurement.
- Analysis of fraudulent activities and profiles of high-risk (legal entity) bidders (NLP, data mining) for correlation with procurement data and alerting high-risk activities.
- Analysis of market conditions (NLP, data mining, econometrics) to identify patterns of corrupt practices while comparing market conditions with offers in submitted bids. Analysis of social media (multi-media analysis) for capturing hints and signs of on corrupt activities.

2.2.3 Pilot 3: Transparent Management of Foreign Aid for Rebuilding Ukraine [led by YC in T5.3]

Ukraine currently confronts Russia's unwarranted aggression and benefits from substantial foreign aid for infrastructure rehabilitation. The EU notably contributes to these endeavours. In January 2023, the Ukrainian State Agency for Reconstruction and Development of Infrastructure (ARU) was established to oversee these initiatives, with approximately €2.75 billion allocated by the government and plans for a sixfold increase. While Ukraine demonstrates political resolve against corruption and utilizes the contemporary Prozorro electronic procurement system since 2016, the absence of supplementary mechanisms for robust oversight by civil society and benefactors, curtails its potential for efficacy.

To realise this pilot, the pilot partners involved will utilise and enrich the following CEDS: Health (procurement processes in healthcare), Industrial (costs of goods), Finance (financial transactions of high-risk legal entities, procurement transactions), Public Administration (procurement data). In this pilot, we will involve, utilise, and enrich the following 50+ public and private datasets in Ukrainian, Russian, Belarus, English, as follows:

- Data from official open national registries (EGRUL/EGRIP portal about legal entities, Russian military lists, PEP lists, sanctions list, media articles, judicial documents, register of corrupt officials, importers / exporters of Ukraine).
- International registries (lists of legal entities in Russia and Europe).
- Data from web (e.g., market places, job sites).
- Social media posts.

The pilot aims to enhance the utilization of governmental and donor data by the Ukrainian authorities, notably the ARU and the European Union (EU), through the integration of contemporary technologies and robust methodologies. CEDAR endeavours to mitigate corruption risks in procurement processes by leveraging diverse datasets,

conducting multifaceted risk assessments of entities and pertinent individuals, including Politically Exposed Persons (PEPs), to scrutinize potential connections with Russia and identify corruption-prone bids. Additionally, advanced data technologies and machine learning (ML) algorithms will be deployed for post-approval project monitoring. Specifically, the focus will be on:

- Analysis of tenders, bids, and bidders (NLP, data mining) for monitoring procurement.
- Multi-factor risk analysis of Ukrainian legal entities and PEPs (using transliteration algorithms suitable for English, Russian, and Ukrainian language) to identify potential anomalies or linkages with Russia.
- Analysis of social media (multi-media analysis) to search for high-risk linkages and to capture hints and signs of on corrupt activities.
- Earth observation of construction sites (CV) to compare contracts and their actual implementation.
- Analysis of cryptocurrency flows (graph-based approaches) for detecting fraud in crypto transactions.

2.3 Types and Formats of Artefacts Generated / Collected

To provide an overview of the different artefacts contributed by the CEDAR partners and identified until M6, including datasets, software, and research items, Table 2 captures their types, a short explanation per artefact and their format template.

Table 2. Artefacts Overview.

#	Artefact Type	Description	Work Package	Format (indicative)	Accessibility
1	Research Item	Algorithms, Models and Meta-models, Policies, Questionnaires, Deliverables, Research Papers	1-4	Binaries / Executables; RESTful APIs; YAML files	Publishable / Non-publishable
2	Software	Code, APIs, microservices, libraries, UI/UX items and dashboard	1-5	Programming languages / frameworks; Code bases / Libraries; Frontend Tools (Figma [13]) and JavaScript Frameworks	Publishable / Non-publishable
3	Dataset	A structured or unstructured collection of data. This data can be raw, numeric, textual, or multimedia. Currently identified Datasets in CEDAR: Synthetic, Open, Raw / Original / Historical, Generated / Enriched /Produced.	5	word or pdf documents, relational databases (MySQL, etc.).xls, .csv, .txt, .json	Publishable / Non-publishable

2.4 Project Artefacts and Access Rights

The different datasets that are currently available and the ones that will be produced in the CEDAR Project, follow the structure presented in Section 2.1. More details about the data type, the related WP number and the abovementioned structure, is thoroughly described through metadata (i.e., without disclosing sensitive information) per Pilot in Appendix A - Dataset List. In a survey contacted during the first months of the CEDAR Project, the input collected from most of the partners is depicted in the following pie chart. The types of CEDAR Artefacts are distributed as 34.25% being datasets, and 65.75% research items, depicted in Figure 1. We clarify here that in the research items we have incorporated the early software artefacts that have been identified by the CEDAR Consortium.

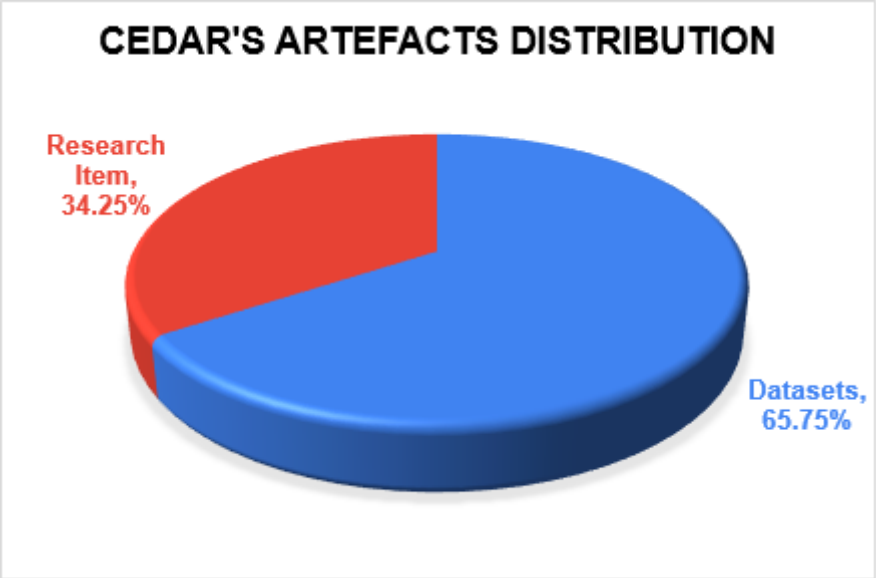

Figure 1. CEDAR Artefacts Distribution.

Table 3 presents the distribution of the artefacts identified within CEDAR by means of numbers and percentages.

Table 3. CEDAR Artefacts Distribution.

Type	Number	Percentage
Research Items	25	34,25%
Datasets	48	65.75%

The following tables present the current status with regards to identified artefacts and their access rights. It is provisioned that those tables are a recurring exercise, and all future updates and additions will be documented under D7.4 and D7.6 Data Management Plan at M18 and M36, respectively.

Table 4. Partners' research item provision to CEDAR.

Identifier	Partner(s)	Artefact Descriptions	Work Package	Task
1	VICOM	Kriptosare	WP4	T4.3
2	VICOM	CorrelaX	WP4	T4.4
3	TRE	NoCorr	WP4	
4	CERTH	Video Captioning	WP4	T4.2
5	CERTH	Object detection	WP4	T4.2
6	CERTH	Concept/event detection, anomaly detection, forecasting	WP4	T4.2
7	CERTH	Image and video manipulation, deep fake detection	WP4	T4.2
8	CERTH	Image and video retrieval	WP4	T4.2
9	CERTH	Fusion of text and visual content	WP4	T4.4
10	KUL	Microsoft Package (Word, Excel, OneDrive)	WP7	T7.3
11	KUL	KUL Library (online and offline)	WP7	T7.3

12	KUL	EU Open Resources (Guidelines, Public Documents…)	WP7	T7.3
13	ICS	Network Intrusion Detection and Reaction	WP3	T3.4
14	ENG	Named Entity Recognition & Relationship Extraction	WP4	T4.1
15	SNEP	Digitalization of procurement process in SBC hospital	WP1	T1.3
16	SNEP	Asset management related to procurement in SBC hospital	WP1	T1.4
17	SNEP	AI services connections	WP3	T3.1
18	SNEP	Cybersecurity for users and dataflow	WP3	T3.4
19	SNEP	Digitalization of data archive in SBC	WP5	T5.2
20	BIGS	Indicators of corruption	WP4	T4.3
21	BIGS	Impact of corruption	WP4	T4.3
22	BIGS	Analysis of results	WP4	T4.4
23	CAU	Data analysis and validation	WP4	T4.4
24	CAU	Relevant variables identification	WP4	T4.3
25	CAU	Impact assessment	WP6	T6.5

Table 5. Partners' dataset provision to CEDAR.

Identifier	Partner(s)	Artefact Descriptions	Publishable / Non-Publishable
ANCE_1	ANCE	Builders (limited company)	Publishable
ANCE_2	ANCE	Builders (unlimited partnership)	Publishable
ANCE_3	ANCE	Workers (Employed)	Publishable
ANCE_4	ANCE	Workers (Self-employed)	Publishable
ANCE_5	ANCE	Work hours	Publishable
ANCE_6	ANCE	Report start of activity	Publishable
INS_1	INS	eProcurement FVG - BUYERS	Publishable
INS_2	INS	eProcurement FVG - RFQ	Publishable
INS_3	INS	eProcurement FVG - RFQ Evaluation	Publishable
INS_4	INS	eProcurement FVG - SELLERS	Publishable
INS_5	INS	GGAP - Construction Planning	Non-publishable
INS_6	INS	GGAP - Service Planning	Non-publishable
INS_7	INS	GIGA - Inspections	Publishable
INS_8	INS	CUC FVG	Publishable
UBI_1	UBI	MDG_Clarity_01	Publishable
UBI_2	UBI	MDG_Clarity_02	Publishable
UBI_3	UBI	MDG_Clarity_03	Publishable
UBI_4	UBI	MDG_Clarity_04	Publishable
SBC_01	SBC	SBC-Tenders	Publishable
SBC_02	SBC	SBC-Bids	Non-publishable
SNEP_01	SNEP	ERAR	Publishable
SNEP_02	SNEP	eNarocanje	Non-publishable
SNEP_03	SNEP	ceneje.si	Publishable
SNEP_04	SNEP	AJPES	Non-publishable
MDP_01	MDP	OPSI	Publishable
MNZ_01	MNZ	Crime meta data	Non-publishable

MNZ_02	MNZ	OSINT e.g. medical devices (icij.org)	Publishable
VOCIM_01	VICOM	Cryptocurrency transactions	Non-publishable
ART_1	ART	ART SM data	Non-publishable
ART_2	ART	ART Scoring data	Non-publishable
YC_1	YC	Ukrainian State registry (https://nais.gov.ua/)	Non-publishable
YC_2	YC	The Register of infringers by NAZK	Non-publishable
YC_3	YC	ANTIMONOPOLY COMMITTEE OF UKRAINE	Non-publishable
YC_4	YC	YouControl Connection analysis	Non-publishable
YC_5	YC	UNIFIED STATE REGISTER OF COURT DECISIONS	Non-publishable
YC_6	YC	International Sanctions List including OFAC, BIS, etc (8 sources)	Non-publishable
YC_7	YC	INFORMATION ABOUT ECONOMIC ENTITIES WITH OUTSTANDING TAX LIABILITIES	Non-publishable
YC_8	YC	SINGLE TAX PAYERS Registry	Non-publishable
YC_9	YC	VAT PAYERS Registry	Non-publishable
YC_10	YC	Large taxpayers registry	Non-publishable
YC_11	YC	Express Analysis by YouControl	Non-publishable
YC_12	YC	Financial Scoring by YouControl	Non-publishable
YC_13	YC	Market Scoring by YouControl	Non-publishable
YC_14	YC	Credit Scoring by YouControl	Non-publishable
BDAP_FVG	CNT	BDAP Expense Reports	Non-publishable
PW_FVG	CNT	Public Works	Non-publishable
TH_FVG	CNT	Territorial Hazards	Non-publishable

2.5 Possible Reuse of the Data

The project will use the data, as presented in Table 5, within the Consortium Members. Within the CEDAR Consortium Access Rights for implementation and exploitation, existing background and general principles have been also defined and signed in the Consortium Agreement (CA) (cf. Section 9; [1]). The same approach will be followed regarding the data published defining the state of the art of the section. If the possibility to improve or exploit results and the technology obtained arises, the data could be used for further specific purposes. This evaluation will be made as the project progresses.

2.6 Expected Size of the Data (if known)

The expected size of the data is unavailable in this initial DMP due to the fact that specific requirements for every CEDAR solution and every hardware specification will be defined in the upcoming period. It is expected that as a research outcome will generate research datasets (i.e., results of the technologies and software, services of the demos, etc.), publications, proposition of new services, dissemination material, etc. The total size will depend on the number of variables used in each use case in the CEDAR Project. Due to size of the project, scope of work and complexity, the expected size of the data cannot be estimated at this moment of the project's lifetime because we foresee data being enriched, collected and generated for both the CEDAR Pilots and the CEDAR Data Space.

2.7 Data Utility

Data and knowledge generated within the project will be used to achieve the project objectives and will be useful in future related research activities. Each dataset will be assessed from the point of view of personal or sensitive data protection; if considered as publicly available. In addition, each dataset will obtain the appropriate permission and

authorization for use by the scientific community and industries working in the field. Data and knowledge generated by CEDAR will be useful to other research projects revolving around CEDAR solutions. Publishable data will be archived in a data-archiving tool such as Zenodo [3] or IEEE Dataport [4] which will allow to make data publicly accessible. CEDAR Consortium aims at generating Open Data so the research community can take advantage of relevant results. All data owners should agree on data openness as far as possible. If the data cannot be made open, the data owners will provide justification. The main issues when considering confidentiality of datasets are: (1) To protect intellectual property regarding underlying business processes, products, and technologies where the data could be used to derive sensitive information that would impact the competitive advantage of the consortium or its members; (2) Commercial agreements as part of the procurements of components or materials that might foresee the confidentiality of data; (3) Personal data that might have been collected in the project; and (4) Due to security constraints, sharing this specific data is prohibited by both national and European legislation.

3 FAIR Data

CEDAR Project supports the reuse of research data and follows FAIR principles [5]. FAIR represents a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable. The international FAIR Principles have been formulated as a set of guidelines for the reuse of research data. The acronym FAIR stands for findable, accessible, interoperable, and reusable. Research data must be of quality that makes them accessible, findable, and reusable. From the beginning of the project the CEDAR Consortium has worked towards making the -currently identified data FAIR, as follows: (1) Findable: data has a unique, persistent ID, located in a searchable resource, and documented with meaningful metadata; (2) Accessible: data is readily and freely retrievable using common methods and protocols, metadata is accessible even if the data is not; (3) Interoperable: data is presented in broadly recognized standard formats, vocabularies, and languages; (4) Reusable: data has clear licenses, and accurate meaningful metadata conformity to relevant community standards and identifying its content and provenance.

Wilkinson et al. [6] were keen to stress that good data management is not intended to be a goal in and of itself, but a means to support continued 'knowledge discovery and innovation'; a further goal could also be considered to support scientific reproducibility and replicability. The Horizon Europe [14] and the European Commission (EC) [7] also advocate for the usage of FAIR principles for data management within its guidelines. Indeed, EC has highlighted the importance of making the data produced by European-funded projects Findable, Accessible, Interoperable and Reusable, with a view to ensuring its sound management, as well as boosting the dissemination of relevant information and the easy exchange of data. Thus, European FAIR data approach implements standards and metadata to make data discoverable, specifying data sharing procedures, under which data may be open, allowing data exchange via open repositories as well as facilitating the reusability of the data. The living documentation for FAIR data is available at GO FAIR [8] but for clarity, the main principles of each pillar will be detailed from subsection Making Data Findable to subsection Making Data Reuseable. In the sections below we set out what FAIR data means in the context of the CEDAR Project by considering these guidelines, the original description from Wilkinson et al. [6], and the GO FAIR Initiative [8]. This deliverable is intended to be a living documentation reflected onto the DMP Register, which will be continuously monitored and kept updated. The DMP Register will be also updated as the requirements and business / user's needs of the CEDAR Project evolve.

3.1 Making Data Findable

CEDAR pays special attention to enhancing the findability and discoverability of the data collected/generated in the course of its activities in accordance with FAIR principles. Storage, processing and sharing (among project participants) will occur via data exchange platforms (such as MS SharePoint [15] repository), whereas interaction with the wider public will be achieved through the official project website (i.e., https://cedar-heu-project.eu/). Data will be anonymized following a privacy-by-design and security-by design approach meaning that data will not identify any individuals and therefore real names of participants or instances within the data will NOT be distributed. Data will be also shared in relation to publications (deliverables and papers) and based on their dissemination level. When this is not seen as being adequate for the comprehension of the raw data, a report will be shared along with the data explaining their meaning and methods of acquisition. The minimum information recorded for data findability is as follows:

- Name of data set: Univocal identifier of the considered data [D_DESCRIPTION_PARTNERNAME_AA]
- Data types: Real-world, historical, or synthetic data in .xls, .csv, .txt, .json, .docx, .pdf.
- Data generation and/or collection: Description of the type of input used to generate the data and the complete methodology and tools used for data collection.
- Purpose: What are the data collected/generated specifically used for?
- Data origin: Where applicable, information from applications to be developed by each partner.

To increase the data searchability, the project utilises the metadata-driven approach. The metadata that will be used to discover the data collected/generated by the project will be suitable to its content and format. To further

reinforce discoverability of the data deemed open, the project will utilise Zenodo [3] and/or IEEE DataPort [4] platform. Zenodo is a free of charge, open data repository, commissioned by EC and created by OpenAIRE [16] and CERN [17] . The service offered by Zenodo can handle any file format of size up to 50GB, enabling researchers, scientists and EU projects to share their results, promoting data reuse within EU but also all over the world. Using Zenodo, it can further increase the discoverability of the gathered data that is going to be made publicly available by CEDAR also by defining and storing additional metadata provided by the uploader. Moreover, it offers the capability, in case the consortium decides so, to restrict data access for a specific group of users (or the public) for a fixed period of time. Zenodo registers and preserves Digital Object Identifiers (DOIs) [18] for all submitted data through DataCite, the leading global non-profit organisation providing identifiers (and specifically DOIs) for research data and other research outputs. The submitted data is preserved using the safe and trusted foundation of CERN's data centre. All these makes the data preserved in Zenodo to be accessible for years to come. With that in mind, for CEDAR's openly available data the metadata standards provided by Zenodo will be used. Metadata for closed data, on the other hand, will follow the metadata standard described in the next sections. IEEE DataPort [4] is an accessible data platform that enables users to store, search, access and manage data. This data platform is designed to accept all formats and sizes of datasets (up to 2TB), and it provides both downloading capabilities and access to datasets in the cloud. It also enables individuals and institutions to: (1) indefinitely store and make datasets easily accessible to a broad set of researchers, engineers and industry; (2) gain access to datasets that can be analysed to advance technology; (3) facilitate data analysis; and (4) reproduce and reutilise research results.

3.1.1 Data Identification Mechanisms

All documents associated with the project will be identified with a project name and unique and persistent document type designator and number that will be given to the coordinator for the submission to EC. Versioning of the document should be part of the document name and title. As per the documents related to project activities and/or deliverables, the tasks or deliverables number will be used to identify the document followed by a brief title of the activity or deliverable.

Example

CEDAR_D7.2-Data Management Plan_v1.0.pdf

3.1.2 Naming Conventions Used and Versioning

Each set of data produced (dataset, deliverables, etc.) will be named in a uniform way and will include a table with a version control. Only documents created by the consortium will be versioned in an incremental manner (i.e., vA.BB where A and B are integer numbers starting from zero). The recommendations to name documents of the project are as follows:

- Choose easily readable identifier names (short and meaningful);
- Not use acronyms that are not widely accepted;
- Not use abbreviations or contractions;
- Avoid language-specific or non-alphanumeric characters;
- Add a two-digit numeric suffix to identify new versions of one document.
- Dates should be included back to front and include the four-digit years: YYYYMMDD.

For deliverables: CEDAR [Deliverable Code]-[Deliverable Title]_vA.BB, e.g.,: CEDAR_D7.2-Data Management Plan_v1.00 (for submission to the Commission).

For datasets: WP [Work Package number] DEMO [Pilot number.pilot activity number]_[description of the activity] i.e.: WP5 DEMO_1.3_Results of demonstration performance.

3.1.3 Standards for metadata creation (if any)

Basic metadata will be used to facilitate the efficient recall and retrieval of information by the project partners and external evaluators and contribute to easily finding the information requested. To this end, all documents related to the project have to include in the front-page, information about author(s) and editor(s), WP, dissemination level and version.

3.2 Making Data Accessible

Datasets collected / generated in the context of the CEDAR Project by default will be made publicly available (FAIR), unless terms and conditions apply that would prohibit this (e.g., IPR, commercial data, sensitive data, etc.). Data from questionnaires to stakeholders will however not be shared in any form but will remain solely and strictly for use within the project. For data that do not need to remain completely closed the Zenodo [3] repository will be used for depositing them. For datasets where restrictions apply in terms of accessing them, Zenodo eases this process of requesting being granted access permission by allowing uploaders of data to present the terms and conditions for access and be notified when a request for access is issued. The minimum information recorded for data accessibility is as follows:

- Accessibility: Publishable / Non-publishable; Open / Confidential.
- Repository: Description / location of the available data.
- Shareability restrictions / related Information: Where applicable, information from applications to be developed by each partner.

As far as project publications are concerned, when appropriate, project findings will be published in journals, press or online relevant fora. According to the GA, for all publications, we will ensure that open access is available, and we will thus aim at publishing through the Open Research Europe (ORE) platform [19]. A high-level strategy will be agreed upon by the partners regarding the content of publications and other communication activities and this will be in line with the Consortium Agreement (CA), which deals with issues related to IP ownership and similar matters. In alignment with the EC Guidelines on Open Access to Scientific Publications and Research Data, as depicted in Figure 2 CEDAR will also follow a combination of Gold and Green Open Access strategy to its scientific publications, which will be agreed upon during the first months of the project execution. Gold Access will be encouraged for high-impact journal publications while the self-archiving, Green Access will be granted for the rest of the publications. The repositories listed in Zenodo funded by OpenAIRE and the repositories available through the consortium members will be considered while there will also be a relevant repository on the website of the project and in social networking sites for scientists and researchers like ORCID [20] and ResearchGate [21].

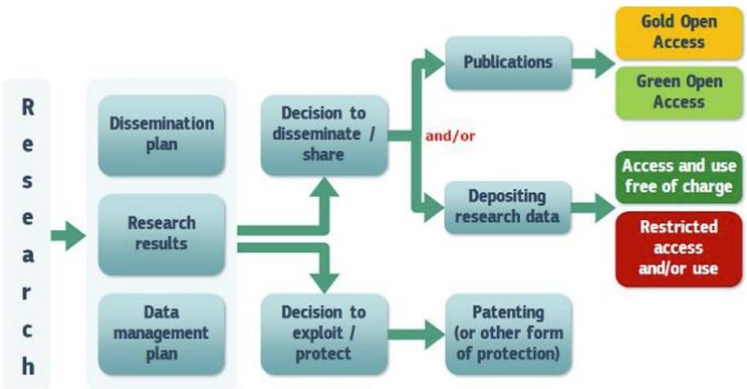

Figure 2. Open Access to Scientific Publication and Research Data [22].

3.3 Making Data Interoperable

Making data interoperable (FAIR) can be achieved by using suitable standards for data and metadata creation while making use of appropriate vocabularies (e.g., for providing search keywords). CEDAR will adopt in its data management methodology (in the datasets where it can be applicable) the use of metadata vocabularies, standards and methods that will increase the interoperability of the data collected/generated through its activities. The assessment of data interoperability will be updated in future reviews to guarantee the CEDAR data and services fit the needs of a specific scenario (such as data / system infrastructures, interests, or purpose of data). The minimum information recorded for data interoperability is as follows:

- Format: Data format, measuring unit, typical order of magnitude, e.g., JSON-like, CSV, etc.
- Expected size of the data: To be defined, 3MB/Day or 2MBday when compressed etc.
- Standards and metadata: The metadata attributes list and the methodologies used.
- Standard Data or Software Interfaces: List of technologies, stacks, code basis, open APIs, data formats for consumers / producers processes, programming languages used to promote results replicability.

The interoperability of the data deemed open will be facilitated through Zenodo, with its metadata being stored internally in JSON format following a defined JSON schema. With regards to the data that will not be publicly shared, CEDAR will adopt the Dublin Core Metadata [23] standard which ensures that the data meet traditional quality and consistency standards while they remain interoperable with other data sources at the same time. The standard includes fifteen metadata elements that introduce a vocabulary of concepts with natural-language definitions that can be converted into machine-readable open formats such as XML, HTML, etc.: hence being processable. The metadata template that CEDAR will use can be found in Table 6.

Table 6.Metadata template for CEDAR datasets.

Identifier	Field	Description
1	Title	A name given to the resource
2	Creator	An entity primarily responsible for making the resource
3	Subject	The topic of the resource
4	Description	e.g., abstract, table of contents, graphics, …
5	Publisher	Only for published items
6	Contributor	Entities that contributed to the making of the resource
7	Date	The termination of the data collection period
8	Type	[dataset, article, questionnaire, …]
9	Format	File format of the resource
10	Identifier	e.g., ISSN if your item has been published
11	Source	Which tools were used to collect the data
12	Language	A language of the resource
13	Relation	A related resource
14	Coverage	The extent or scope of the content of the resource.
15	Rights	Information about rights held in and over the resource

3.4 Making Data Reuseable

Reusability (FAIR) of data is the last aspect of the FAIR data principles discussed in the present document. To provide for the reuse of data, the datasets will be accompanied by a relevant license. CEDAR considers the family of Creative Commons Licenses (CCL) (https://creativecommons.org/licenses/) as a very straightforward way to allow the reuse of data as they ensure that the source and authority of the data are recognized and commercial interests -if applicable- can also be protected. The specific version of the CCL license (or any other license - if different for some reason) used is dataset-dependent. The minimum information recorded for data reusability is as follows:

- **Reuse of existing data:** Specifications as follows: No reuse of existing data; generation of synthetic datasets based on their original characteristics; reuse of existing, historical, real-world data in logs, etc.
- **Data backup:** Consistent location of the data, including previous releases.
- **Quality Consistency:** Constraints determining the quality of the collected data.
- **Simulation / Synthetic data generation tools:** Description / location of possible method or synthetic data generation model useful for generating new data.

4 Other Research Outputs

At this stage of the project, other relevant research outputs have been identified as research item artefacts including models, software, libraries, etc. As described in Sections 2.3 and 2.4, the types of CEDAR artefacts are distributed as 34.25% being of research item artefact type. The distribution is depicted in Figure 1, while Table 4 enumerates the research items. In the upcoming period, the CEDAR Consortium will also consider which of the aspects pertaining to FAIR data above, can apply to the management of other research outputs, and how they will be managed and shared, or made available for re-use, in line with the FAIR principles.

5 Allocation of Resources

Each entity is responsible for managing their data. CERTH as project coordinator, UBI as Data Manager and UPM as leader of the Task 7.2 CEDAR Scientific, Technical, and Data Management, will supervise that data management. Besides, all partners are responsible for data generation, metadata production and data quality and they will have specific responsibilities depending on the data and the internal organization in the WPs and tasks where data is created or used.

As far as the costs required for managing the data collected / generated during CEDAR's activities FAIR in accordance with the FAIR Principles, those are included within the budget of the project. These estimated costs will be needed to cover a set of specific data processing and data management activities, spanning from collection and documentation through storage and preservation over to sharing and reuse. These activities are part of the WP under which the respective data are processed so the required effort will be part of the respective WP. Each work package leader is also responsible to follow-up data management in their WP. In the first instance, each entity is responsible for managing their data. Therefore, each partner is responsible to collect or generate data and metadata in an adequate way to the standards of Horizon Europe. Specific responsibility is to be assigned depending on the data and the internal organization in the WP and tasks where such data is created.

The cost needed for making the research data compliant with the Horizon Europe requirements are eligible under CEDAR Project, following the obligations of article 17 and Annex V of the Grant Agreement. As the management of the data following this DMP is aligned with the commitments of the CEDAR Project, the cost for making the data compliant with FAIR principles can be assumed by the corresponding budget inside the project for each partner involved. Task 7.2 of the Grant Agreement, called CEDAR Scientific, Technical, and Data Management, will run all over the project, being the specific task that covers the practical implementation of the DMP. Using Zenodo (free of charge) ensures that costs for long-term preservation of the data are negligible. In any case, a better view on the costs will be available in a later stage of the project.

6 Data Security

As mentioned earlier under Chapter 2, CEDAR consortium will handle the related research data in accordance with the principles of privacy by design and security by design. Any gathered data will be securely handled throughout the entire duration of CEDAR to protect it from loss and unauthorized access. Personal or sensitive data (e.g., from tabular instances or images) will be anonymized and be only accessible to those who are authorized to access it. CEDAR adopts privacy-by-design and security-by-design principles.

Most specifically, the partners will store their data in the project common repository held in Microsoft Teams, administrated by CERTH (as described in D1.1 Project Handbook) and also in the private servers and data-storage systems of each consortium partner. All partners responsible for processing data[3] have the responsibility to implement the appropriate technical and organizational measures concerning the security of processing, hence ensuring that the data remains protected under all necessary security controls (including backup policies and integrity checks performed every week[4]) and access controls (identification, authentication, authorization) within their infrastructure. In this respect, each partner will act at first sight as an independent data controller (article 24 of the GDPR) in terms of complying with the respective obligations dictated under the GDPR including those associated with the security of the data used for the development of the project.

Article 11 of the GDPR states that if the purposes for which a data controller processes personal data do not or do no longer require the identification of a data subject, the controller shall not maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with GDPR. In the unfortunate event of personal data breach, the project partners will notify without delay their competent national supervisory authorities as well as the data subject(s) that may be affected by the breach. At the same time, they will document any personal data breaches and all related information. Indeed, according to the Section 10 Non-disclosure of information of the CA 'All information in whatever form or mode of communication, which is disclosed by a Party (the "Disclosing Party") to any other Party (the "Recipient") in connection with the Project during its implementation, and during its term, that has been marked as "confidential" at the time of disclosure or, when disclosed orally, has been identified as confidential, at the time of disclosure and has been confirmed and designated in writing within 15 (fifteen) calendar days from oral disclosure at the latest, as confidential information by the Disclosing Party, is "Confidential Information".' For data to be retained exclusively in the respective partners cloud environments, no additional security is expected to be needed from Coordinator's perspective. Every partner is responsible for ensuring that the data are stored safely and securely and in full compliance with legal regulations such as GDPR. After the completion of the project, all the responsibilities concerning data recovery and secure storage remains to the partner storing the dataset.

In case additional safeguards are needed, the project will use secure cloud services, access control and authentication mechanisms that include the security settings of the specific cloud hosting providers. These systems include advanced key management systems, authorization procedures and encryption options, as well as specific control and visibility for complying with the legal framework. Regarding open data, security for long-term preservation relies on the, widely tested, Zenodo platform [3].

[3] Processing, according to Regulation (EU) 2016/679 of the European Parliament (GDPR), means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

[4] The integrity check is the process of comparing the current state of stored data and/or programs to a previously recorded state to determine any alteration or change.

7 Ethics

The CEDAR Consortium is aware of the ethical aspects pertinent to the scope of CEDAR Project, which are addressed under the Work Package 7 on T7.3 "Ethical and Legal Analysis". In particular, T7.3 puts particular emphasis on two aspects pertinent to the scope of CEDAR: (1) the involvement of research participants; and (2) the protection of the personal data to be collected and further processed for the purpose of CEDAR Project. T7.3, thus, focuses on the Demonstrators' activities by capturing, for instance, the procedures for onboarding research participants and by producing an overview of the technical and organizational measures for the protection of personal data of the research participants. Additionally, consortium partners are committed to comply with the ethical principles as set out in article 14 of the GA and their Annex V. All the activities must be carried out in compliance with the European Code of Conduct for Research Integrity of ALLEA [24]. Data that will integrate the DMP will be considered as non-personal data, applying Regulation EU 2018/1807, allowing that this information could be processed freely throughout the EU. In this way, all the cautions to exclude the application of personal data regulations, mainly GDPR, will be adopted. Article 15.2 of the GA regulates the processing of personal data by the beneficiaries establishing that beneficiaries must grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement. For example, to allow interaction with individuals in dissemination acts (e.g. workshops, social media, newsletter…) partners will be compliant with the article 5.1.c of the GDPR (data minimization). This principle states that the data to be gathered by the partners in the implementation of the project must be:

- adequate – sufficient to properly fulfil your stated purpose;
- relevant – possess a rational link to that purpose;
- limited to what is necessary – you do not hold more than you need for that purpose.

The main objective of the foreseen research activity is to test, assess, and evaluate the proposed solution and its effectiveness. Any personal data derived (name, surname, e-mail address, image, video, audio, faces, footages, etc.) will abide to the data minimization principle, ensuring that there will be well-rounded and sufficient explanation on the reasons that urge the project to collect the said personal data per case and for the purpose of the project. After the end of the demonstration, all personal data will be immediately transferred to encrypted and/or secure and password protected servers or devices. Before data can be used, they will be anonymised, unless there is an explicit agreement with the research participant that says otherwise, e.g., in the case of photos. The transcripts will delete/modify any information that would enable you to be identified (names, locations, etc.) directly, by inference or by association. This Anonymisation will be complete and irreversible as the original audios will be destroyed.

Furthermore, the Consortium will be careful in the way they formulate and publish their research findings to avoid the stigmatization or stereotyping of any of the involved groups. No environmental damage, political or financial adverse consequences and misuse are foreseen as potential outcomes of the CEDAR activities. Due to the human involvement and personal data collection, the project will comply with all the relevant legislation, e.g., the General Data Protection Regulation (Regulation (EU) 2016/679), the free movement of such data, the Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), as well as all the specific national legislations on the protection of personal data. Informed consents and information sheets including specific information on data management and participants' personal data protection rights will be developed during the lifespan of the project, adjusted to the needs of the foreseen activities, and communicated in understandable language and terms.

In addition, CEDAR will draft an Incidental Findings policy, describing all potential unexpected findings not being related to the purpose of the research and concerning disclosure of info around the commission of any unlawful act critical for the safety and the security, that subsequently require immediate disclosure. Dedicated statements informing the participants involved in the research activities regarding incidental finding policy along with specific

examples per country will be included in the provided information sheet. Each participant of dissemination & communication activities will be considered as data subject, having the right to exercise control over their personal data, determining the extent to which it can be gathered/re-used and eventually processed (compliant with chapter III of the GDPR). Furthermore, our website will be fully compliant with the GDPR including a legal notice, privacy policy statement and cookies warning. With these provisions visitors of the website will be aware about our use of their personal data in the project. No additional information about the informed consent procedures for data sharing and long-term preservation regarding the use of personal data, can be provided now (month 6), but it will be included in D7.4 (M18), and D7.6 (M36) dealing with other data related issues, requesting to provide an explanation of the human skill assessment, presumably to be done in an automated way, and raising some issues of human assessment by algorithms. We clarify that ethics or legal related issues will be reported in D7.3, D7.5 and D7.7 Ethical and Legal Assessment at M12, M18 and M36, respectively.

8 Conclusions

This deliverable reports the activities and methodology adopted by the CEDAR Consortium towards efficient and transparent Data Management within the project until M6. Our activities have focused on enumerating and keeping the collected, used and generated datasets updated to promote transparency, reproducibility, and collaboration within the CEDAR Project and the scientific community in general. By establishing the DMP Register, we have ensured that data is organized, preserved, and FAIR, both for the current project and for future endeavours.

In the next period, we plan to keep the DMP Register up-to-date and enrich it with additional datasets as the CEDAR Project evolves.

9 References

[1] CEDAR Consortium Agreement, version 1.0
[2] HE Data Management Plan Template. Available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan_he_en.docx
[3] Zenodo – Research. Shared. Available at: https://zenodo.org/
[4] IEEEDataPort at: https://ieee-dataport.org/
[5] FAIR Data Guide. Available at: https://horizoneuropencpportal.eu/repository/5b7fcc0e-73da-4e76-8b46-3682a36fa59b
[6] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data3:160018 doi: 10.1038/sdata.2016.18 (2016).
[7] European Commission, "H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020," November 2016. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [Accessed May 2024].
[8] "FAIR Principles," GO FAIR, 2019. [Online]. Available: https://www.go-fair.org/fair-principles/ [Accessed May 2024].
[9] European Commission. (2023). Grant Agreement: Project 101135577 - CEDAR. Directorate-General for Communications Networks, Content and Technology, CNECT.G – Data, G.1 – Data Policy and Innovation. Associated with document Ref. Ares (2023)7141900 - 20/10/2023
[10] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[11] https://gdpr-info.eu/
[12] https://blog.unguess.io/what-is-security-by-design-the-bestapproach-to-cybersecurity [Accessed February 2023]
[13] https://www.figma.com/
[14] https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm
[15] https://www.microsoft.com/en/microsoft-365/sharepoint/collaboration
[16] https://www.openaire.eu/
[17] https://home.cern/
[18] https://en.wikipedia.org/wiki/Digital_object_identifier
[19] https://open-research-europe.ec.europa.eu/
[20] https://orcid.org/
[21] https://www.researchgate.net/
[22] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm
[23] https://dublincore.org/
[24] European Code of Conduct for Research Integrity of ALLEA (All European Academies). https://allea.org/code-of-conduct/ [Accessed June 2024]

10 Appendix A – Dataset List

Dataset Information Template (completed by each partner in charge of the dataset).

10.1 Pilot 1 – Transparent Management of National RRP Funds in Italy

Number of datasets identified	9
Partner	INS
Dataset description	The e-procurement platform by Insiel supports entities (buyers and sellers) with roles-specific access, featuring attachments in various formats (PDF, P7M, DOCX, XLSX) for information fulfillment. It includes data on RFI/RFQ activities, evaluation criteria, attachments like blueprints, quotations, and evaluation reports from committees. The platform also covers three-year public works planning (construction and services) for all FVG region buyers, detailing projects, bids, and statuses. Additionally, it provides data on construction site conditions, worker presence, equipment, external suppliers, non-conformities, and information on framework agreements and related service contracts.
Related WP/Task	WP5 / Task 5.1
Will you reuse any existing data? If YES, how?	The Dataset is currently used by all public authorities of region FVG.
Methodologies for data collection / generation	The Dataset has been collected from a backup of the current database of the e-Appalti FVG platform
Data type(s)	Text
Data format(s)	DB SQL(mandatory) + Attachments(optional)
Data storage	INS data center
Expected size of the data	>> 280 GB
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	Data Scientist interested in evaluating algorithms and methologies aimed at identifying anomalies (e.g.: criminal activities, corruption, etc,)
Data access	Data is private and owned by INS and FVG but might be licenced
Data privacy	Yes, data need to be treated carefully with a high level of security during transmission and analysis.

Number of datasets identified	6
Partner	ANCE
Dataset description	The dataset provides information about builders' companies and workers in the building sector in the FVG Region. It includes details on the number of hours worked monthly by each worker and the start dates of building activities. This data aims to track company and worker activities within the regional construction industry.
Related WP/Task	WP5 / Task 5.1
Will you reuse any existing data? If YES, how?	The Dataset is currently used by Casse Edili and ANCE.

Methodologies for data collection / generation	The Dataset has been collected from a backup of the current database of Casse Edili.
Data type(s)	Text
Data format(s)	FILE EXCEL
Data storage	Casse Edili
Expected size of the data	<<30 GB
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	Data Scientist interested in evaluating algorithms and methologies aimed at identifying anomalies (e.g.: criminal activities, corruption, etc,)
Data access	Data is private and owned by INS and FVG but might be licenced
Data privacy	Yes, data need to be treated carefully with a high level of security during transmission and analysis.

Number of datasets identified	3
Partner	CNT
Dataset description	The data includes expense reports, public works details, and information on potential risks and hazards (geological, hydraulic, forest fires, seismic zones, avalanche sites) for several municipalities in the FVG Region.
Related WP/Task	WP5 / Task 5.1
Will you reuse any existing data? If YES, how?	The Dataset is currently used by Casse Edili and ANCE.
Methodologies for data collection / generation	The Dataset has been collected from a backup of the current database of Casse Edili.
Data type(s)	Text
Data format(s)	CSV/API
Data storage	FVG
Expected size of the data	<<15 GB
Metadata and standards	Yes.
For whom might the dataset be useful?	Public Authorities; Regulatory Authorities; Data Scientists
Data access	Italian Open Data Licence
Data privacy	There is no data privacy since it is open data

10.2 Pilot 2 – Transparent Management of Slovenian Public Healthcare Funds

Number of datasets identified	2
Partner	SBC
Dataset description	The report includes data on past tenders from 2023 and 2024, alongside data on submitted bids to these tenders, providing comprehensive insights into recent procurement trends and bidding behaviours.

Related WP/Task	WP5 / Task 5.2
Will you reuse any existing data? If YES, how?	The data is currently used by SBC
Methodologies for data collection / generation	Gathered by SBC over the years when managing tenders and bids.
Data type(s)	Text
Data format(s)	Word documents, PDFs, Excel files
Data storage	SBC
Expected size of the data	TBD
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	End users for monitoring procurement process. Data Scientist interested in evaluating algorithms and methologies aimed at identifying anomalies (e.g.: criminal activities, corruption, etc,).
Data access	TBD
Data privacy	SBC private data. To be opened.

Number of datasets identified	4
Partner	SNEP
Dataset description	The data includes public money use in Slovenia, covering individual transactions, public procurement e-invoices, and payments. It details procurement between €10-40k, item prices on the Slovenian market, and the Slovenian business register.
Related WP/Task	WP5 / Task 5.2
Will you reuse any existing data? If YES, how?	The data is public and reuseable.
Methodologies for data collection / generation	Gathered via API, and FTP, CSV2SQL transformer.
Data type(s)	Text
Data format(s)	JSON, CSV
Data storage	https://erar.si/ https://www.enarocanje.si/#/english
Expected size of the data	>> 10 GB
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	End users for monitoring procurement process. Data Scientist interested in evaluating algorithms and methologies aimed at identifying anomalies (e.g.: criminal activities, corruption, etc,).
Data access	Public data under licence
Data privacy	Private data. Exploring opportunities for collaboration.

Number of datasets identified	3
Partner	MDP, MNZ
Dataset description	An Open Data Portal provides public access to data about past corrupt or fraudulent activities, promoting transparency and accountability by enabling citizens and researchers to analyse and understand historical instances of misconduct.
Related WP/Task	WP5 / Task 5.2
Will you reuse any existing data? If YES, how?	The data is reuseable.
Methodologies for data collection / generation	Gathered via different ways.
Data type(s)	Text and different kind of data
Data format(s)	CSV and other
Data storage	https://podatki.gov.si/data/search?open_data=True, and MNZ
Expected size of the data	TBD
Metadata and standards	DCAT-AP
For whom might the dataset be useful?	End users for monitoring procurement process. Data Scientist interested in evaluating algorithms and methologies aimed at identifying anomalies (e.g.: criminal activities, corruption, etc,).
Data access	Data can be licensed with CC BY 4.0 or TBD for MNZ Data.
Data privacy	Published under Open Data Directive, and Private data.

10.3 Pilot 3 - Transparent Management of Foreign Aid for Rebuilding Ukraine

Number of datasets identified	2
Partner	ART
Dataset description	Public data from social media related to Ukraine and metrics data can reveal potential connections between tender participants, suspicious mentions, and other risk factors, helping to identify and mitigate fraudulent activities.
Related WP/Task	WP5 / Task 5.3
Will you reuse any existing data? If YES, how?	The Data is currently used by Artellence and TBD.
Methodologies for data collection / generation	Generated based on ART SM data by applying ML algorithms
Data type(s)	Different data types (numerical, categorical, textual)
Data format(s)	API
Data storage	Artellence Data Hub
Expected size of the data	TBD
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	Public Authorities; Regulatory Authorities; Data Scientists

Data access	National Public data
Data privacy	N/A

Number of datasets identified	14
Partner	YC
Dataset description	The Unified State Register includes AMCU's decisions on anticompetitive actions like bid rigging by financial entities. Our algorithm identifies company or person connections using official Ukrainian court data and aggregates the International Sanctions List for comprehensive analysis.
Related WP/Task	WP5 / Task 5.3
Will you reuse any existing data? If YES, how?	The Data is currently used by YouControl.
Methodologies for data collection / generation	The data is being updated a few times per day.
Data type(s)	Different data types (numerical, categorical, textual)
Data format(s)	API
Data storage	YouControl DataHub
Expected size of the data	API requests
Metadata and standards	No particular metadata or standards have been followed while collecting the dataset
For whom might the dataset be useful?	Public Authorities; Regulatory Authorities; Data Scientists
Data access	Public data
Data privacy	Open data

Number of datasets identified	1
Partner	VICOM
Dataset description	Cryptotransactions stored in the blockchain (mainly Bitcoin)
Related WP/Task	WP5 / Task 5.3
Will you reuse any existing data? If YES, how?	The dataset is currently used in many other project
Methodologies for data collection / generation	The data has been downloaded.
Data type(s)	Text/Raw data
Data format(s)	API
Data storage	Local storage of VICOM
Expected size of the data	2 TB
Metadata and standards	Could contain OSINT information regarding the entities scraped from the clear web
For whom might the dataset be useful?	End user for monitoring transactions and detect possible fraud or money laundering operations.

Data access	Public data
Data privacy	N/A

10.4 Other Datasets

Number of datasets identified	4
Partner	UBI
Dataset description	Open Data provides expense information for Greece's Ministries of Health and Social Security, Infrastructure and Transportation, Climate Crisis and Citizens Protection, and Defense.
Related WP/Task	WP2 / Task 2.1
Will you reuse any existing data? If YES, how?	The Dataset is currently used by all public Greek authorities for transparency.
Methodologies for data collection / generation	The dataset will be connected to the public OpenDataAPI (https://diavgeia.gov.gr/api/help) and analytics will be made available to the CEDAR Data Space
Data type(s)	Text
Data format(s)	JSON to Elasticsearch
Data storage	Local storage of Ministry of Digital Governance (GR)
Expected size of the data	<<20 GB
Metadata and standards	JSON / XML
For whom might the dataset be useful?	Public Authorities; Regulatory Authorities; Data Scientists.
Data access	Public data
Data privacy	GDPR Compliance

11 Appendix B – DMP Register

The metadata and attributes of the DMP Register (as of the 31st of May 2024) that are maintained in the CEDAR's project repository are depicted in Figure 3.

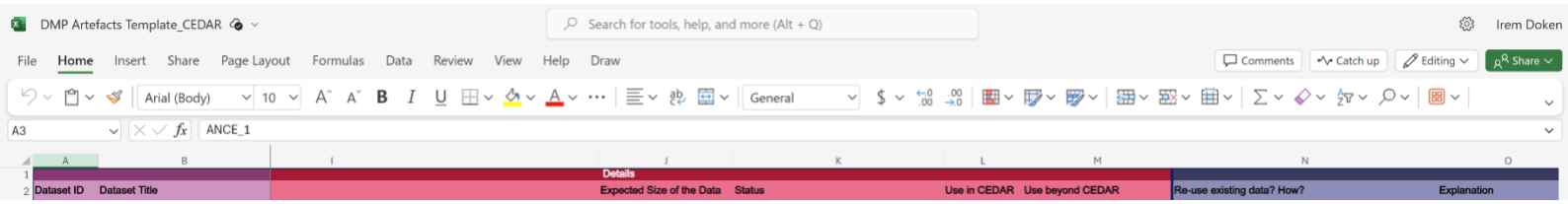

Figure 3. DMP Register.