TRUSTAI

TRANSPARENT, RELIABLE
& UNBIASED SMART TOOL

D8.7 Data Management Plan V2

September 30, 2022

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant agreement No 952060.

DOCUMENT CONTROL PAGE

DOCUMENT	**D8.7 – Data Management Plan V2**
TYPE	**Report**
DISTRIBUTION LEVEL	Public
DUE DELIVERY DATE	30/09/2022
DATE OF DELIVERY	30/09/2022
VERSION	0.3
DELIVERABLE RESPONSIBLE	INESC TEC
AUTHOR (S)	Fábio Moreira (INESC TEC), Gonçalo Figueira (INESC TEC), Yulia Karimova (INESC TEC), João Aguiar Castro (INESC TEC)
OFFICIAL REVIEWER/s	Gonçalo Figueira (INESC TEC)

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Yulia Karimova (INESC TEC)	09/06/2022	First draft
0.2	Fábio Moreira (INESC TEC)	15/09/2022	Second draft considering partners' contributions. Version submitted to reviewers.
0.3	Fábio Moreira (INESC TEC), Gonçalo Figueira (INESC TEC)	30/09/2022	Final revision
0.4	Fábio Moreira (INESC TEC), Gonçalo Figueira (INESC TEC), João Aguiar Castro (INESC TEC)	25/01/2023	Revised version (after review meeting).

DISCLAIMER:

The sole responsibility for the content lies with the authors. It does not necessarily reflect the opinion of the CNECT or the European Commission (EC). CNECT or the EC are not responsible for any use that may be made of the information contained therein.

Table of Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
CC	Creative Commons
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
EU	European Union
HCXAI	Human-centered Explainable AI
KPI	Key Performance Indicators
MS	Milestones
PM	Person Month
PR	Press Release
SMEs	Small and Medium-sized Enterprises
WP	Work Package
XAI	Explainable Artificial Intelligence

1. General information

Creators of DMP	• Yulia Karimova (INESC TEC/ FEUP) • Gonçalo Reis Figueira (INESC TEC) • Pedro Amorim (FEUP / INESC TEC) • Fábio Neves Moreira (INESC TEC) • Cristiane Ferreira (INESC TEC) • João Aguiar Castro (INESC TEC)
Affiliation	INESC TEC
Funder	European Union’s Horizon 2020 Research and Innovation program under the Grant Agreement (GA)
Template	Horizon 2020
ORCID	• Yulia Karimova (https://orcid.org/0000-0002-1015-6709) • Gonçalo Figueira (https://orcid.org/0000-0001-6696-824X) • Pedro Amorim (https://orcid.org/0000-0001-7857-3405) • Fábio Moreira (https://orcid.org/0000-0003-4296-8207) • Critiane Ferreira (https://orcid.org/0000-0002-8680-1938) • João Aguiar Castro (https://orcid.org/0000-0002-5475-5156)
Grant Number	952060
Project duration	01.10.2020 – 30.03.2025 (4.5 Years)

1.1. Project abstract

Artificial intelligence is single-handedly changing decision-making at different levels and sectors in often unpredictable and uncontrolled ways. Due to their black-box nature, existing models are difficult to interpret and hence trust. Explainable AI is an emergent field, but, to ensure no loss of predictive power, many of the proposed approaches just build local explanators on top of powerful black-box models. To change this paradigm and create an equally powerful, yet fully explainable model, we need to be able to learn its structure. However, searching for both structure and parameters is extremely challenging. Moreover, there is the risk that the necessary variables and operators are not provided to the algorithm, which leads to more complex and less general models. State-of-the-art, yet practical, real-world solutions cannot come only from the computer science world. Our approach, therefore, consists in involving human intelligence in the discovery process, resulting in AI and humans working in concert to find better solutions (i.e., models that are effective, comprehensible, and generalizable). This is made possible by employing 'explainable-by-design' symbolic models and learning algorithms, and by adopting a human-centric, 'guided empirical' learning process that integrates cognition, machine learning, and human-machine interaction, ultimately resulting in a Transparent, Reliable and Unbiased Smart Tool. This proposal aims to design TRUST, ensure its adequacy to tackle predictive and prescriptive problems and create an innovation ecosystem around it, whereby academia and companies can further exploit it, independently or in collaboration. The proposed 'human-guided symbolic learning' should be the next 'go-to paradigm' for a wide range of sectors, where human agency/accountability is essential. These include healthcare, retail, energy, banking, insurance, and public administration (of which the first three are explored in this project).

1.2. DMP Objectives and Scope

The DMP is a working document that specify the general data management practices for the project. Among others aspects, it will provide an ongoing assessment of the datasets the use-cases will generate over time. In the current version, the DMP is particularly focused in good practices to enable the FAIRness of data, such as the selection of repositories services and the overall metadata to represent the project datasets. The DMP encompasses a set of baseline activities that will be specified in each version of the DMP as more concrete strategies are adopted according to the data requirements.

The DMP will be updated every six months by the INESC TEC´s data steward, João Aguiar Castro, hereafter referred to as DMP manager. Compliance with the DMP will be assessed by the DMP manager together with the project coordinators.

2. Data summary

In this project, we aim at learning a directly human-comprehensible model. We focus on machine learning applications that use images, text data, and tabular data, including classification/regression. We want to improve the performance-explainability trade-off, so ML would be to combine the performance of (deep) neural networks and the explainability of theory-based approaches. We will mainly use approaches considering symbolic learning methods that work with analytical expressions that humans can understand and improve on. These methods will support a human-centric system, using theory and insight from humans to guide the machine's empirical search. Humans are a key element of this project, as they need to interpret the output of the system and guide the empirical search for machine learning. The resulting tool is TRUST – a Transparent, Reliable and Unbiased Smart Tool, that can be used in practice in a vast number of applications and disrupt multiple sectors, where human control is essential. To achieve our vision, we propose a disruptive, but viable, paradigm with the objective to reach a breakthrough, rather than incremental improvements. To that end, we mobilize a multidisciplinary consortium (detailed information about consortium in section 3 of this DMP), joining experts from computer science, neuro and behavioral science, and industrial engineering, which have recognized international leadership in their own areas.

The objectives of this project are balanced between:

I. the development of the novel paradigm,
II. its exploration in different application domains, and
III. the creation of an innovation ecosystem around the TRUST human-centric framework.

Moreover, we integrate three key areas:

1. Machine learning, to search and provide explainable expressions, with as few task-specific hard-coded constraints as possible, and to allow the user to easily tweak the explainability;
2. Cognition, to incorporate behavioral and cognitive models of human explainability and causal reasoning in providing contrastive explanations (why P rather than Q?) and connecting causes;
3. Human-machine interaction, to establish trust and promote cooperation, and to interactively adjust the explainability of the ML model and its solutions, depending on the user's level of expertise.

The project has three real-world use cases with different types of data, methodology, and management rules. Each use case will start by providing a toy problem, with core features, and then be progressively extended, as human experts validate the models and explanations generated. This DMP includes all necessary related Research Data Management (RDM) issues about each case separately. Moreover, this project includes 5 EU countries: Portugal, Estonia, France, Netherlands, Cyprus, and 1 non-EU country: Turkey. Thus, this DMP also includes detailed information about data management related to the Turkish partner (which can be found in section 3 of this DMP).

2.1. Use cases

2.1.1. Cancer treatment (Healthcare)

This use case includes collaboration between Estonia, France, The Netherlands, Portugal, and Turkey, where the Dutch entity, CWI, is responsible. The use case will be developed at Leiden University Medical Center (LUMC). The use case is about paraganglioma. For this type of tumor, which is usually benign and slow growing, it is uncertain whether patients will need treatment to remove the tumor in the future. Normally, treatment is postponed until a persistent growth of the tumor is demonstrated, or the tumor starts to cause severe symptoms. The goal of this use case is to forecast the development of paraganglioma to make better decisions about the moment of treatment and follow-up. In order to support the physician's decision at each meeting with the patient, we plan to set the prediction tasks to be quantifiable and observable outcomes. Input to the prediction is all the information collected up to that time about the patient. The main challenge here is to forecast whether the patient will need treatment and if so, when. The key to this is likely being able to predict the growth of the paraganglioma. In addition, forecasting the probability of possible negative effects of the tumor such as hearing loss can help in better pinpointing the moment to proceed to treatment. The main users of the models will be clinicians. However, since there is shared decision-making between the clinicians and patients, the clinician needs to be able to explain these models to the patient. The use case aims at delivering the desired predictions through explainable AI models that clinicians can accept, validate, and use in clinical practice. To develop TRUST approaches, several data types will be used, including images, tumor-related data, DNA mutations, biochemical screenings, past treatments, and other generic features (e.g., sex, medication). The data is available after all the consent procedures have been undertaken according to the hospital policies and after the approval of the Medical Ethical committee. The raw data collected by the hospital during the time of the project is highly sensitive and due to privacy reasons it cannot be shared outside the hospital. Only processed data that ensures anonymity (e.g., patients cannot be identified) can potentially be shared, provided that a data transfer agreement has been set in place, typically after careful consideration and negotiation between the legal departments of the parties in question. Currently, data is only accessible at LUMC and experiments on the data are carried on at the LUMC. Since CWI already has its management processes in place to deal with hospital partnerships, a data transfer agreement between CWI and LUMC is under consideration by the legal departments of the said institutes. This could enable speed ups thanks to the high-computing facilities available at CWI.

At the current phase of the project, the tabular data is collected and data available from before the project is being used. The collection of the image data has been started but is not finalized yet.

If use case 1 will involve research towards clinical aspects (specifically: in research where the software is used as a medical device and its use is studied prospectively on patients), the following documents/information must be created: (i) Final version of study protocol as submitted to regulators/ethics committee(s), (ii) Registration number of clinical study in a WHO-or ICMJE- approved registry (with the possibility to post results), (iii) Approvals (ethics committees and national competent authority if applicable) required for invitation/enrolment of the first subject in at least one clinical center. Copies of pertinent opinions/approvals by ethics committees and/or competent authorities for the research with humans also will be created and kept on file. The outcome of future medical ethical committee submissions will be shared, and, as is common practice at LUMC, be kept on file. More details about informed consent can be found in the Grant Agreement. The responsibles for data management of these data are Peter Bosman at CWI and Tanja Alderliesten at LUMC and at CWI.

Table 1 - Summary of healthcare data characteristics

Use Case	Healthcare
Type of data	Tabular data and images
Data source	LUMC
Responsible partners	CWI
Brief description	The data include several measurements of metrics describing the paraganglioma tumors related to several patients. These metrics, measured at several points in time, englobe features related to the size and volume of

	the tumor. Additionally the data includes patient information such as DNA mutation, biochemical screenings, past treatments, and other generic features (e.g., sex, use of medication).
Purpose	Track patient data and tumor evolution
Privacy level	Confidential

2.1.2. Time slot selection (Online retail)

This use case includes collaboration between four countries, Estonia, France, Portugal, and Turkey. Three Portuguese entities, INESC TEC, Sonae MC, and LTPlabs are involved, where the latter is responsible. This use case is focused on online retail in different industries, including fashion, electronics, and grocery. More concretely, it is focused on decisions to be made on the fly (i.e., in real time), such as selecting delivery time slots to be offered, from where items will ship, and by what shipping method. Logically, AI systems are necessary to make these decisions in an optimized way and these decisions need to be based on and simulated using historical data, in order to achieve and assess the fitness of each expression. During this use case, the data that will be used is generally not sensitive and not personal. The most sensitive information could be the composition of the shopping basket and the address location. However, customers' privacy will be ensured by their anonymization. The first type of data is related to customers (demographic data, shopping history, current basket, destination, customer profile, purchasing habits) and it will help to predict how much each customer is willing to pay to have his/her order arrive within a specified delivery window and to extract insight on customer preferences. The second type of data concerns logistics (logistics load, planned routing, delivery conditions, logistics costs) that will help to accurately estimate the cost of serving a customer within a given time window and to the ability to provide a cost breakdown. In this use case, we will involve a large grocery retailer, Sonae MC, which has more than 200 brick-and-mortar stores, and a considerable volume of online orders (and which has committed to providing all the necessary raw data). We focus on the time slot selection problem since it has a huge impact on the optimization of last-mile deliveries. In addition, when selecting time slots, there is a trade-off to be considered between customer satisfaction (i.e., offering the most appropriate slots for a given customer profile) and operational efficiency (i.e., aggregate orders in certain slots to optimize transportation). This process also involves multiple stakeholders in the company, namely marketing, and operations divisions, though no data involving these stakeholders is necessary. Raw data concerning retail customers are sensitive, personal, and private so they cannot be opened publicly, but the publishing of the processed data (anonymized, codified, or pseudonymized) will be decided during the project, probably in the middle of the project. Other partners should not need the raw data related to this use case. In case any partner needs realistic data to develop algorithmic approaches or realistic cases to test the prospective interfaces, the data will be anonymized and masked so that no connection exists to a real person or entity. The responsible for data management of these data is André Morim (LTPlabs). The ownership of these data belongs to Sonae MC. The raw data will never be shared, only the realistic samples generated from anonymized data can be shared or published (i.e., as part of a scientific publication). The type of data that can possibly be published will be defined upon the date of the refereed publication.

Table 2 - Summary of online retail data characteristics

Use Case	Online retail
Type of data	Tabular data
Data source	Sonae
Responsible partners	LTP
Brief description	The data include information about past customer orders, detailing important features such as the shopping basket value, delivery time slot prices, chosen time slots, and customer location.
Purpose	Track customer online order data to solve a dynamic time slot pricing problem
Privacy level	Project consortium

2.1.3. Demand forecast (Energy)

This use case includes collaboration between Cyprus, France, The Netherlands, Portugal, and Turkey, where the Cypriot entity, Apintech, is responsible. There are two subcases considered; one where forecasting addresses the building domain and the hourly forecasting resolution and one where the country level forecasting is addressed with a yearly resolution.

This former case is about reading and processing, in real-time, the data from all required data sources, related to the energy in the building, in a trustful way to train a forecasting ML model and use it subsequently in daily, hourly, and minute demand prediction. In this TRUST use case, we will focus on a single building (that can provide essential real-time energy/ indoor quality data). Data about non-linear factors (weather, indoor conditions of a building, behavioral data) will be collected to help with the performance and accuracy of electrical energy consumption forecasting, using, as examples alone, NN/ LSTM, STS and GP methods. Moreover, additional data related to the building as well as location weather data (temperature, dry bulb temperature, dew point temperature, wet point temperature, air temperature, humidity, wind speed, wind direction, brightness of sun, precipitation, vapor pressure, global/solar radiation, sky condition); indoor data (ambient temperature, electric usage, occupancy) will be collected. This data will be used on the development of a suitable and open data API (available at htps://ds.leiminte.com). Note that no personal or sensible data is involved in the open API. Again, the consumer's privacy will be ensured, by anonymization/pseudonymization, using the best practices. Since the necessity to transfer data to other partners is present, anonymized samples will be available to be shared. The idea is to provide TRUST partners with realistic data so that XAI algorithms and interfaces can be developed with a sufficient degree of realism. All the datasets shared will ensure that no connection between real persons or entities can be established. Good quality datasets will be critical to the main objective of this use case which is to develop an advanced and explainable energy forecasting tool for city building and also for entire cities. Furthermore, each energy consumption forecast is supposed to be disaggregated into a more detailed form, increasing the granularity of the forecasts, and guiding the users by providing causal rules and counterfactuals that can be used to support decision-making. Multiple interfaces are to be developed for each type of user (e.g., building owners and policymakers).

As to the country level subcase, this involves data macroeconomic data that are publicly available and there is no associated privacy issue involved.

The updated information on energy data sources and sharing, regarding this second version of the DMP, is described in the following points:

- Building real-time data will be uploaded via file uploads and most importantly via an open and public TRUST-AI API; any third party may use this API to upload data. Accessing data from cloud storage (Google, Dropbox, etc.) will also be considered. In the case of TRUST-AI and for the testing requirements the provider of this data will be APINTECH. A separate business model will be considered for this service as we consider it is significant.
- Weather data will be uploaded via an open and public API; the provider will be the openweathermap weather service. The API will access the vast amounts of data offered by this provider, under his set terms.
- Socio-economics data will be uploaded via file uploads. The ODYSSEE MURE database will be used for demonstration purposes. This provider currently offers no API access and supports only file downloads. As ODYSSEE is an ongoing initiative, one will need to follow its developments regarding any data access policy changes it may introduce.
- TRUST-AI application-specific data will be uploaded via a private API. In TRUST AI these will refer to behavior and pricing data. However, there is no limitation as to what can be accommodated here. The provider of this data for the TRUST-AI project will be APINTECH. This data will need to observe the same schema as public data. Its privacy relates to its non-public visibility alone and not to its structure.

Table 3 - Summary of the energy data characteristics

Use Case	Energy
Type of data	Tabular data
Data source	APINTECH

Responsible partners	APINTECH
Brief description	The data include energy consumption and indoor comfort metrics at several granularity levels.
Purpose	Track and predict energy consumption in different buildings, as well as at country level.
Privacy level	Project consortium

2.2. Data Inventory Register

In order to gather information about the data collected or produced by the project, a Data Inventory Register (DIR) was developed, and it is available to all TRUST members in the INESC TEC cloud drive service.

The DIR can be defined as a comprehensive catalogue of metadata, wherein information about the datasets can be incrementally updated. The DMP manager is the person responsible for maintaining the DIR, while the completion of the information for each dataset is the responsibility of the researchers involved in the collection and use of each dataset, or by a designated person in each use case.

The information provided must be as rich as possible, and metadata fields must be filled in where applicable. The following metadata elements, depicted in table 4, are included in the DIR, and some of the information collected can then be used when sharing the datasets. The information collected in the DIR will also be used to update the Data Summary section of this DMP, as the DMP will be used to keep track of the TRUST data.

Table 4 – Data Inventory Register metadata elements

Metadata element	**Description**
Use case	The identification of the use case
Responsible Party	The person(s) responsible for the creation/maintenance of the dataset. It can be a point of contact for the dataset.
Dataset Name	A descriptive relative short name for the dataset, with contextual information – *what, where, when.*
Data Type	Example: experimental, observational or simulation data. Data from surveys, climate, biophysical data, etc...
Personal Data	If the dataset contains information relating to an identified or identifiable natural person. Yes or No.
Sensitive Data	Any data that reveals a subject´s information. Namely, racial or ethnic origin, political and religious beliefs, biometric data, sexual orientation and others. Yes or No.
Date Created	Date of creation of the dataset. Example: YYYY-MM-DD. It can be a temporal coverage: YYYY-Mx – YYYY-Mz.
Number of Data Files	The total number of files that make up the dataset.

Source	Where the data originates. Is the data reused? Recommended best practice is to identify the related resource.
Format	The file format. Example: txt/xml/csv.
Size	The total weight of the dataset (when applicable) Example: 4gb
Dataset Duration	(When applicable) The extant or time taken to play or execute the dataset. For instance, the duration of a video file.
Data Collection Instrument	The instruments used to generate and or process the data.
Support documentation	Whether documents have been created to provide context for the data. For exemple, experimental protocols and readme files.
Dataset Duration	The frequency with which the data will be updated.
Data Collection Instrument	Place where the data is stored. For instance, institutional drive.
Update frequency	The frequency with which the data will be updated.
Storage Location	Place where the data is stored. For instance, institutional drive.
Preservation	Backup periodicity and for how long the dataset needs to be preserved.
Availability	Private, consortium or open
Rights and Restrictions	Data licence. CC BY-SA 4.0 is recommended for data sharing.

2.3. TRUST outputs

All the outputs of the project are to be preserved under internal and trustable storage platforms (emails, DPIA, Consent informs, agreements, Communication & Dissemination Plan (CDP), NDA non-disclosure agreements, research code, research papers). Part of the data and outputs of the project can be useful to reuse in different sectors such as healthcare, retail and energy, banking, insurance, and public administration since decision-making in these contexts must be transparent and explainable. In other words, the data is very important to markets with great societal and economic potential. Moreover, the development, implementation, and exploitation of TRUST in those sectors must be supported by a thorough discussion of the organizational, societal, ethical, and legal implications that these tools might have. This should also contribute to the wider debate on AI systems, decision support (e.g., medical prescription, operations management, financial investments), decisions that affect people (e.g., whether customers or applicants are being treated fairly), and accountability providence to external entities (e.g., regulatory bodies, citizens).

3. *FAIR* data

In this section, the guiding principles for promoting the FAIRness of the TRUST data are outlined. It describes high-level practices on how the project intends to make the data *findable*, *accessible*, *interoperable* and *reusable*. However, the availability of data has to be considered on a case-by-case basis, considering the requirements that different types of data may have.

The following principles are considered to enable the FAIRness of data:

• For the datasets that reach the publication stage, persistent identifiers will be assigned to the data and included in its metadata.

• Metadata should be accessible even when the data may not be publicly available due to data protection issues.

• Metadata is based on standard vocabulary(ies), whenever possible.

Table 5, provides the projected availability for both raw and processed data by use case. This information will be further detailed as the DMP is updated. However, eventhough there are several limitations, the project consortium will make an effort to make data publicly available as much as possible. These data publications will be evaluated on a case-by-case basis, in order to comply with the FAIR principles.

Table 5 – Datasets expected availability (ongoing update)

Use Case	Raw	Processed
Cancer treatment (Healthcare)	Cannot be shared given that the data is sensitive	Derived data may be published following NDA
Time slot selection (Online retail)	Cannot be shared as they belong to Sonae MC	Sample datasets subject to anonymization may be shared with consortium members
Demand Forecast (Energy)	Anonymized data shared in the public domain	Anonymized data shared in the public domain

When publishing project results, TRUST will consider the following elements in the Data Availability Statement, aligned with possible publications' data policies.

Table 6 – Data Availability Statement

Availability status	Statement elements
Open	• Data repository or platform, where the data can be accessed; • DOI • Title of the dataset • Licence • List of data items with description of contents
Restricted	• Ethical, legal or commercial reason the data cannot be shared openly • Link to permanent record detailing restrictions and conditions for access (data deposit under private visibility)
Third party	• Citation

3.1. Repository selection for data availability

As a rule of thumb, the data generated by the project, provided it does not contain sensitive information, will be deposited, and preserved, in the data repository provided by INESC TEC (https://rdm.inesctec.pt/)

The INESC TEC institutional data repository, based on the CKAN open source data management system, can be accessed by any user at any time and from any location. Therefore, the data will be available for both project members and the general public. The data backups are performed daily, while tape backup is performed weekly.

The publication of data is coordinated by the DMP manager and the project coordinators, or directly with the person(s) responsible for the creation of the dataset.

Not all data needs to be deposited for sharing and preservation purposes. Therefore, the datasets to be deposited at INESC TEC are those that can be made publicly available. Moreover, datasets can be deposited with «private» visibility, with respect to embargo periods, while only the dataset metadata is made public. The conditions for accessing data that contain some kind of restriction will have to be evaluated in each case, while also taking into account the INESC TEC´s Data Protection Officer recommendations.

The deposit of data in disciplinary repositories is also open to comply with possible data policy requirements from publishers. If no disciplinary repository is recommended by the publisher, the selection of a disciplinary repository, to comply with more restrictive policies, will be assessed via the data repositories directory re3data.org (http://re3data.org/), with the support of the DMP manager.

As a complementary deposit approach, Zenodo (https://zenodo.org), the catch-all repository for EC funded research, will be considered as the service to support the dissemination of project achievements. These outputs can take the form of processed data, research protocols, reports, as well as the different version of the DMP itself, among other artefacts.

Finally, significant data and models related to the Energy Use Case (country level) have been shared at the Open Science Foundation, OSF (www.osf.io). As the data has been posted there more than a year ago and is referenced as such in the literature, they will necessarily remain also at this repository.

3.2. Licencing

TRUST will aim to made datasets available under Creative Commons (CC). CC licences are well suitable for research data due to their conformity to both copyright law and database rights and are applicable in all jurisdictions. By default, the license for TRUST data sharing is Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) (https://creativecommons.org/licenses/by-sa/4.0), which specifies that:

- Credit must be given to the creator.
- Adaptations must be shared under the same terms.

As an alternative, CC Zero allow end-users freedom to use and re-use data without restriction, and is also a licence advocated for sharing research data. More restrictive licences can be adopted when applicable.

3.3. Data Identifier

In order to make Trust data FAIR, a unique identifier, such as the Digital Object Identifier (DOI), needs to be assigned when the data, or its metadata, is publicly available, providing a stable and consistent way to locate both the data and metadata.

The DOIs for the data that will be deposited in the INESC TEC data repository will be registered and managed via the DataCite Fabrica (https://doi.datacite.org/), a service provided by the leading provider of research data DOIs. By the time a dataset reaches the publication stage, the DMP manager will trigger the process of minting a DOI, by filling out a set of metadata associated with the dataset, which must include the minimal citation elements, described in the next sub-section.

Complementarily, outputs directly deposited in other repositories will be granted a DOI generated by the service, or according to the repository policies.

3.4. Data Citation

Regarding data citation, the following minimal data citation elements are recommended, according to DataCite:

- Creator (PublicationYear): Title, Publisher. DOI

The dataset title must indicate the subject matter, the geography and the time period covered by the data, when applicable.

Additional properties may also be added: Version (in the case of a dynamic database it should be used the date of download); Resource Type (e.g. dataset, database, table, map, sound file, image).

3.5. Data Documentation

Data sharing within and outside project must include metadata, along with the needed documentation for others to understand and reuse the data. Data documentation is a very important process throughout the entire data life-cycle, as it can be helpful for those who created the data as well as for those with whom the data is shared. TRUST data may need to be documented at various levels, and a ***readme.txt*** file should be adopted as a common practice to accompany the datasets, being included in a common folder, at the time of sharing the datasets.
Some information may already be integrated in the software (from which the data is derived) or platforms (where the data will be made available), namely the metadata provided by the INESC TEC data repository (see Table 7).

Table 7 – Data Documentation elements

Level	Information elements
Dataset	• Technical report for the user to understand how the data were collected and processed • Citation information to indicate how secondary users will have to cite the data. • Study design, methodology, instruments and measures used.
File or database	• Information on how the files that make up the dataset (or tables in a database) relate to each other. • The format of the files.
Variable and item level	• Information on how the object of analysis was produced. • Label explaining the meaning of variables.

3.6. File Name Convention

Correct file naming not only makes it easy to access data, but also to understand what a data file is and its content. File names have to remain meaningful and useful beyond their original creation and storage – it should outlast the person who created the file. There are several elements that must be consider in the implementation of the file name strategy, particularly when sharing data with others. On its own, the filename should convey metadata about the data file.

Given the diversity of the data that TRUST will generate, specific aspects of the file naming strategy will be specified in the following updates of the DMP. However, the following good practices will be considered when naming the files:

Context information: Data files name have to include specific or descriptive information.

Dates in the Year-Month-Day format: to maintain the chronological order and simplify the process of sorting and browsing data files. Hence, TRUST will comply with the international standard for the representation of date and time, ISO (https://www.iso.org/iso-8601-date-and-time-format.html).

Consistency: The adopted convention has to be followed systematically.

Short and relevant names: By norm, 25 characters is enough to convey descriptive information.

Use of underscore or hyphen over full-stops or spaces: The latter are parsed differently on different operation systems.

File name quick access guide: A simple guide must be displayed in the project drive to ensure that others can easily understand the file naming adopted for different types of data, and to protect against data loss.

3.7. Metadata standards

The INESC TEC data repository already follows an appropriate metadata standard, specifically Dublin Core, a highly adopted, domain-agnostic metadata standard, published as ISO Standard 15836 :2017 (https://www.iso.org/standard/71339.html). Metadata records based on the Dublin Core standard ensure that project data is more easily findable and also interoperable.

Additional custom metadata fields can be added in the INESC TEC data repository, to address specific metadata requirements. For the deposit of a dataset in the INESC TEC data repository a metadata template will be made available, by the DMP manager, in the project drive. This template is to be filled in by the data responsible before the deposit. Table 8 details the core metadata elements in the INESC TEC data repository.

Table 8 – Metadata template for dataset publication

Metadata element	Content
Visibility	Public or private.
Title	The title of the dataset
Tags - subject	Topics related to the dataset
Licence	Creative Commons Attribution Share-Alike as default
Version	The version of the dataset
Author	The person(s) responsible for the dataset
Author email	The contact of the responsible for the dataset
Maintainer	The person responsible for the management of the dataset
Contributor	An entity responsible for making contributions to the dataset
Spatial Coverage	Spatial characteristics of the dataset
Temporal Coverage	Temporal characteristics of the dataset
Type of Instrument	Type of instrument used for data collection or capture
Creation Date	Date of creation of the resource YYYY-MM-DD
Format	The file format, physical medium, or dimensions of the resource
Publisher	An entity responsible for making the resource available
Relation	A related resource. Recommended practice is to identify the related resource by means of a URI. A string may be used conforming to a formal identification system
Type	Dataset as default
Source	The source from which the dataset is derived
File Size	The total volume of the dataset (how much storage it consumes). It can also provide information regarding the duration of the dataset
Software	Computer program used to generate and run the data

4. Allocation of resources

4.1. Costs for making the data *FAIR*

To make the data FAIR it is necessary to have access to the chosen research data repositories. The INESC TEC repository, and Zenodo, are free and open-access, thus no additional costs are required. The processed data that will be stored at INESC TEC cloud drive is also freely accessible to all the partners of the project. The scientific publications resulting from TRUST will also be shared at INESC TEC drive. Recall that, these publications are to be preferably published in open-access journals but if that is not the case, public versions should be provided through INESC TEC drive links.

In sum, during the project the following assets will be used:

- Hardware/devices: work desktop, laptop computers, and institutional servers (costs of this equipment are detailed in the Grant Agreement document).
- Software: Windows, Linux, Office 365, Visual Studio.
- Cloud services: https://drive.inesctec.pt, with access granted to all partners through provided links of shared folders. Maintenance activities related to INESC TEC drive are to be carried out by the INESC IT structure.

The aforementioned information and procedures are to be maintained in this version of the DMP.

4.2. Responsibilities for data management in the project

A multidisciplinary Consortium was created with the participation of experts from R&D organizations (INESC TEC, INRIA, NOW-I, CITIS), universities (University of Tartu, LUMC), and ambitious SMEs with different profiles: Apintech works with real-time, big data collected from sensors and IoT, LTP conducts analytics-based consulting, and Tazi AI commercializes an explainable AI platform. The organizational structure of the Consortium includes an independent Ethics and Data Protection Board (EDPB), composed of elements of the different institutions, and necessarily including those involved in the use cases (INESC TEC/LTPlabs, CWI/LUMC and POLIS21). EDPB includes a Data Protection Officer from each WP leader and pilot:

- UC1: LUMC data protection officer, infoavg@lumc.nl
- UC2: Vasco Dias, dpo@inesctec.pt
- UC3: Costas Daskalakis, info@apintech.com

This consortium body will monitor the ethics issues involved in this project and how they are handled. Additional responsibilities include:

- Responsible for DMP creation:
  Gonçalo Reis Figueira (https://orcid.org/0000-0001-6696-824X),
  Fábio Neves Moreira (https://orcid.org/0000-0003-4296-8207),
  Yulia Karimova (http://orcid.org/0000-0002-1015-6709),
  João Aguiar Castro (https://orcid.org/0000-0002-5475-5156).
- Responsible for DPIA, Grant Agreement, and other legal documents. creation: Vasco Dias (INESC TEC)

Responsible for the RDM in each use case:

1. **Cancer treatment (Healthcare):**

- Responsible for the collection of the data: Peter Bosman (CWI)
- Responsible for the processing and preservation of the data: Peter Bosman (CWI)
- Responsible for backups: Peter Bosman (CWI)
- Responsible for publishing and sharing data: Peter Bosman (CWI)

1. **Time slot selection (Online Retail):**

- Responsible for the collection of the data: André Morim (LTPlabs)
- Responsible for the processing and preservation of the data: André Morim (LTPlabs)
- Responsible for backups: André Morim (LTPlabs)
- Responsible for publishing and sharing data: André Morim (LTPlabs)

2. **Demand forecast (Energy):**

- Responsible for the collection of the data: Nikos Sakkas (Apintech)
- Responsible for the processing and preservation of the data: Nikos Sakkas (Apintech)
- Responsible for backups: Nikos Sakkas (Apintech)
- Responsible for publishing and sharing data: Nikos Sakkas (Apintech)

4.3. Potential value of long-term preservation

The data acquired, collected, and generated during this project are unique and thus very important for other researchers as well as for educational purposes. Some data will be preserved to enable initially unforeseen uses of the data and to guarantee fully documented and reproducible data from the project, ensuring the reuse of the data in multiple domains and sectors such as healthcare, retail and energy, banking, insurance, and public administration, and different applications.

Some data will be preserved at the Zenodo repository without any limitation after the project without any embargo period (more detail will be described in the following DMP versions). The INESC TEC IT structure will be responsible for any action related to the long-time preservation in accordance with the repository guarantees. The corresponding information will be added, if necessary, in a future version of the DMP. Each preserved dataset of the project will have Digital Object Identifiers (DOIs) attributed by the repository (Zenodo or INESC TEC).

The Use Case leaders have decided not to share any data through repositories until now, but this is still a possibility to evaluate after or when the first research papers are published.

5. Data security, access, storage and backups

To prevent against the risk of data loss, data must be stored in institutional network drives (intranet or partners 'networks), who must be routinely backed up and provide account authentication systems to prevent unauthorized access, preferably by resorting to strong passwords. It is highly recommended to store data in centrally managed network drives, thus ensuring the storage of data in a single place. Moreover, data must be available when needed via VPN.

In TRUST-AI, different criteria have also been defined according to each use case:

1. **Cancer treatment (Healthcare):**

All pertinent ethics committee opinions/authorizations were submitted as a report by the EDPB, before the beginning of work in use case 1. The raw data never will pass to other partners from LUMC. The measures to protect the data are following good clinical practices. Moreover, LUMC anonymizes the data before sharing it with CWI, where the data are processed. Minimizing the risk of identifying any individuals as much as possible.

In this use case, the raw data will be preserved in the database of the LUMC hospital, and it will not be shared with others. Only anonymized data will be shared with others through the INESC TEC cloud service available at https://drive.inesctec.pt. This concerns the data typically also shared in a research paper. The backups of the processed data will be managed by the INESC TEC drive maintainers. These backups are usually preserved for more than one year. All the data that is available at the INESC TEC drive will also be accessible to all partners of the TRUST-AI project. To access and process data, licensed software is used to reduce the risk of intruders. Moreover, all the computers have properly updated antivirus software. The browsers have Adblock installed to block pop-up pages and other insecure connections. All technical issues related to the software will be controlled by each member of the project and in case of necessary support, they contact the IT staff of the responsible entity. The paper documents are not to be preserved unless they are indispensable to continue the project or to serve as proof of any kind.

Therefore, the procedures adopted in the first version of the DMP are maintained in the second version.

2. **Time slot selection (Online Retail):**

In this use case, the raw data provided by Sonae MC is preserved on a database owned by LTPlabs is not shared with others. Only small samples of processed data are shared with others through the INESC TEC cloud service available at https://drive.inesctec.pt. The data that is being shared depends on the necessities of each of the involved partners, yet those data are always processed (anonymized and masked) so that no connection between real persons or entities can be defined. The raw data have periodic backups according to the internal policies of LTPlabs. The backups of the processed data are managed by the INESC TEC drive maintainers. These backups are usually preserved for more than one year. All the data that is available at the INESC TEC drive is also accessible to all partners of the TRUST-AI project.

To access and process data, licensed software is used to reduce the risk of intruders. Moreover, all the computers have properly updated antivirus software. The browsers have Adblock installed that permit blocking the pop-up pages and other insecure connections. All technical issues related to the software will be controlled by each member of the project and in case of necessary support, they contact the IT staff of the responsible entity. The paper documents are not to be preserved unless they are indispensable to continue the project or to serve as proof of any kind.

The procedures adopted in the first version of the DMP are maintained in the second version.

3. **Demand forecast (Energy):**

Data is shared through an open API (in real-time) and repository available at https://ds.leiminte.com and through the INESC TEC cloud service available at https://drive.inesctec.pt

All the data that is available at the INESC TEC drive will also be accessible to all partners of the TRUST AI project. To access and process data, licensed software is used to reduce the risk of intruders. Moreover, all the computers have properly updated antivirus software. The browsers have Adblock installed to block pop-up pages and other insecure connections. All technical issues related to the software will be controlled by each member of the project and in case of necessary support, they contact the IT staff of the responsible entity. The paper documents are not to be preserved unless they are indispensable to continue the project or to serve as proof of any kind.

According to the aforementioned information regarding the third use case, the procedures adopted in the first version of the DMP are maintained in the second version.

Furthermore, from a global project perspective, the raw data will not be transferred from a non-EU country to the EU, nor in the case related to Turkey. If any processed data transfer occurs, all appropriate documentation will be provided and described in a new version of the DMP.

All technical and organizational measures will be described on DPIA according to art. 35º GDPR. In further versions of the DMP, this information will also be added. Moreover, it will be described the security measures that will be implemented to prevent unauthorized access to personal data, or the equipment used for processing, anonymization/pseudonymization techniques, and data protection policies.

More detail about the right access can be read in the Consortium and Grant agreements.

6. Ethical aspects

The TRUST-AI Project is financed by Horizon 2020 (Nº 952060) and responds to all existing requirements related to the Research Data Management and Protection of Personal Data. All partners follow the Regulations defined by the EU and other agreements defined by the Consortium of this project which define a set of rules to guide projects. This second version of the DMP will be verified by the INESC TEC DPO, will follow the General Data Protection Regulation, and will be monitored every year to add all changes that occurred in the project.

TRUST-AI will adopt a data protection by design and default approach. The general principles are described in the following sub-sections.

6.1. Data Minimization

Following a key principle in data protection law, TRUST-AI will only collect essential personal information data, and this data should only be kept for the time necessary to carry out the purposes for which it was collected, or the time required to comply with applicable law.

Hence, personal data must be reviewed periodically to decide whether unnecessary identifying information is retained.

When identifying information is no longer needed it will be safely deleted or destroyed. Conventional file deletion is not enough to ensure that the data cannot be recovered. A suitable strategy to securely delete data will be considered, following recommended practices.

6.2. Data Anonymization

When considering the availability of sensitive data, without revealing the confidential information it contains, data anonymization procedures must be carried out. As soon as the identifying information is no longer needed, direct identifiers should be removed, where possible, by deleting them or replacing them with pseudonyms. Namely:

- First and last name
- ID
- Email address
- Phone number
- Physical features, voices, photos or other images

To perform a pseudo-anonymization of the personal and sensitive data, collected data will be treated with a code. Thus, linkage files, containing information to link data subjects to identifiable individuals, must be encrypted and stored securely and separately from the de-identified data.

Particular attention should be paid to indirect identifiers, such as age, education and employment information, that can still make it possible to identify subjects. In this case, advanced anonymization techniques will be considered.

For datasets reaching the publication stage, only the dataset metadata will be made publicly available, and a point of contact will be designated for requesting access to the data.