D4.3: TREASURE platform data lake (1st version)

30/06/2022 (Project Month 13)

Author: TXT e-solutions S.p.A. (TXT)

Technical References

Project Acronym	TREASURE
Project Title	leading the TRansion of the European Automotive SUpply chain towards a circulaR future
Project Coordinator	POLITECNICO DI MILANO (POLIMI)
Project Duration	36 months as of 1 June 2021

Deliverable No.	D4.1
Dissemination level 1	CO
Work Package	WP4 “TREASURE platform design, development & integration”
Task	T4.2 “TREASURE data lake development”
Lead beneficiary	TXT
Contributing beneficiary(ies)	UNIZAR, SUPSI, MARAS, EDGE, SEAT
Due date of deliverable	30.06.2022
Actual submission date	30.06.2022

Document history		
V	Date	Beneficiary partner(s)
V0.1	25/05/2022	TXT
V0.2	10/06/2022	TXT
V0.5	20/06/2022	(Internal Reviewers)
VF1.0	30/06/2022	TXT

1PU= Public

PP= Restricted to other programme participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

DISCLAIMER OF WARRANTIES

This document has been prepared by TREASURE project partners as an account of work carried out within the framework of the EC-GA contract no 101003587. Neither Project Coordinator, nor any signatory party of TREASURE Project Consortium Agreement, nor any person acting on behalf of any of them:

a. makes any warranty or representation whatsoever, express or implied,
    i. with respect to the use of any information, apparatus, method, process, or similar item disclosed in this document, including merchantability and fitness for a particular purpose, or
    ii. that such use does not infringe on or interfere with privately owned rights, including any party's intellectual property, or
    iii. that this document is suitable to any particular user's circumstance; or
b. assumes responsibility for any damages or other liability whatsoever (including any consequential damages, even if Project Coordinator or any representative of a signatory party of the TREASURE Project Consortium Agreement, has been advised of the possibility of such damages) resulting from your selection or use of this document or any information, apparatus, method, process, or similar item disclosed in this document.

EXECUTIVE SUMMARY

The present deliverable D4.3 “TREASURE platform data lake (1st version)” is the first document released within T4.2 “TREASURE data lake development”. The document accompanying the software release of the preliminary version of the platform centralized storage location where all the data relevant for the TREASURE Platform resides. The document complements the description of platform design provided in D4.1, focusing on the Data Layer that consists in the sub-architecture assigned to contain all the components that heavily leverage data flow and manipulation. The Data Layer represents the repository on which the Service Layer (the second sub-architecture) is based for the implementation and integration of the main components and tools.

More in detail, the Data Layer consists in two main sub-components: the Data Lake per se which collects relevant information from several data sources and the Data Importer integrating the acquired knowledge within the platform.

- The Data Lake includes data gathered from external databases (IDIS, MISS, Insurances DB); logs and feedback of disassembly procedures performed through WEAVR: Augmented Reality; semantic information and sentiment analysis provided by SSNA tool and information shared by BoL actors on car parts/components material and value; environmental KPIs and life cycle information released by the sustainability tool and recycling KPIs and recycling/disassembly routes determined by the recycling simulation tool. The data deriving from these different items is integrated in the platform as well as the information coming from Circular Economy data experts, to expand the knowledge on best recycling practices in the automotive industry.
- The Data Importer is an accessory component with the purpose of integrating data coming from external sources into the unified TREASURE Platform. It provides ingestion capabilities that allows to connect to many different databases leveraging various technologies, and convert the information gathered into the appropriate format suitable for the Data Lake architecture.

Table of Contents

Table of Figures

1 Introduction

Deliverable D4.3 starts from the outcomes and considerations achieved from the discussion with the TREASURE use cases that lead to the design of platform architecture, which has been reported into D4.1 submitted at M10. The main goal of this deliverable is to accompany the release of the basis for the Data Layer sub-architecture that defines the data source and key tools assigned to information ingestion within the platform.

An overview of the underlying technical components and main actors involved is exposed in Chapter 2, following the architecture overall description which is presented to illustrate the Data Layer role and dependency with other key elements of the platform. Chapter 3 and 4 are focused on the detailed depiction of the two main elements that constitutes the data storage, that is the Data Lake and the Data Importer. Firstly, the component scope and functionalities are outlined to summarising the general information about the software to be released. Then, specific data concerning internal architecture, technological stack, technical manual for installation and licensing (including third parties' components) are reported. A special focus is dedicated to the description of the different type of data source that are exploited by the Data Lake, detailing the data structure, the type and format of data store (relational/non-relational data store, publicly accessible or private, on premise or cloud-based) and, if applicable, the technologies used and datatypes samples. Finally, the conclusions and future steps for the components implementation and integration close the chapter.

1.1 Project Overview

TREASURE – "*leading the TRansition of the European Automotive SUpply chain towards a circulaR futurE"* wants to support the transition of the automotive sector towards Circular Economy (CE), by providing a concrete demonstration of how the industry can benefit from the adoption of Circular Economy practices and principles, both from a business and a technological perspective. One of the main encountered issues highlighted by the automotive actors, refers to the huge information gap existent between Beginning-of-Life (BoL) and End-of-Life (EoL) actors along the whole automotive value chain up to the final consumers.

TREASURE aims at filling this gap through the development of an AI-based assessment tool able to connect and facilitate the interaction among the key involved stakeholders dedicated to car electronics: car parts suppliers, car makers, dismantlers, and shredders. On the other hand, TREASURE goal consists in assisting both BoL and EoL actors in performing their operations, (best recycling options for optimal recovery), taking the most suitable decision according to up-to-date information, as well as in assessing the impact and the effect of their decision on the final customers.

To this aim, a web-based platform will be developed as a new information sharing tool among all stakeholders, both in forward and backward directions, ensuring secure access and confidentiality. The platform will indeed be developed in order to enhance the connection among the actors, making information available through specific modules that will be built and tailored according to their needs.

The platform will be tested with a set of dedicated demonstration actions within the project boundaries. However, it will be designed assuring that its potential can go beyond the project and its sustainability will be properly defined and agreed with the TREASURE consortium, guaranteeing the possibility for its scale-up and adoption by a wider group of stakeholders.

1.2 Scope of the deliverable

This document is the first release concerning T4.2 activities. Therefore, D4.3 has to be considered as a specification of the sub-architecture related to the Data Layer for the actual design, development and integration of the TREASURE Data Lake. Since a continuous iteration and validation process is foreseen, the Data Lake structure will be refined in the following months in order to establish the final version which will be described in D4.4, due on M30.

1.3 Contribution to other WPs

The present document is the first step of the implementation of the WP4 “TREASURE platform design, development and integration” provided as input in D4.1. Since the Data Layer is part of the platform architecture, it is evident that this deliverable lays the foundation of the technical execution of WP5 activities related to platform application, testing and validation in selected uses cases with the aim at reconfiguring the disassembly and recovery process. The TREASURE Platform will then be validated in the demonstration phase performed within WP6, evaluating the new procedure performances in terms of circularity and economic feasibility. Moreover, the present deliverable also represents the starting point for Advisory Tool implementation including its functionality, leveraging on the methodological framework discussed and presented in D2.2. Technical validation and verification will be executed in Task 4.6.

2 Overview of Software Components

This chapter briefly presents the overall architecture of the TREASURE Platform, providing a birds-eye view of the involved components. Then, a more detailed description is given for the components that are more relevant in the scope of this deliverable, explaining the main functionalities exploited by each one of them. Finally, a section will be dedicated to the actors' characterization in which it is highlighted the role covered and the interaction model with the presented components.

2.1 Architecture overview

With respect to the general architecture presented in D4.1 and reported in figure 1, the TREASURE Platform is split into three sub-components, each one containing a set of tools that represents a different level of abstraction on the technical architecture itself.

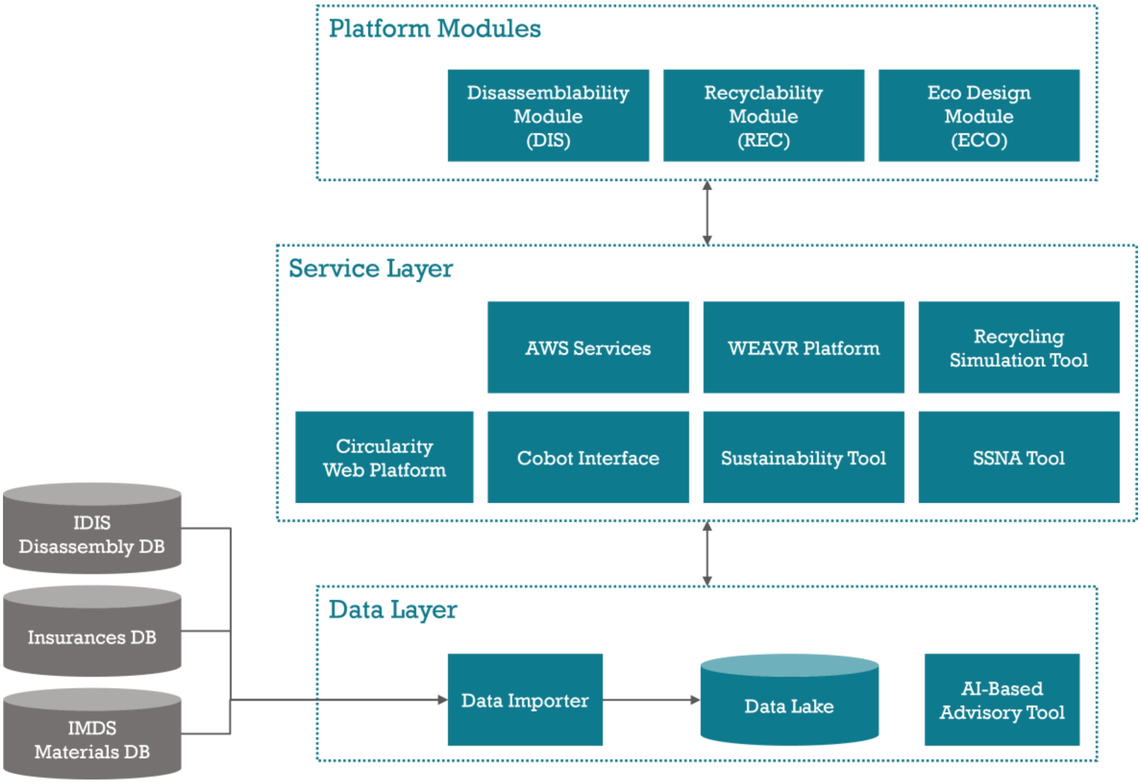

Figure 1 Overview of TREASURE technical architecture

At the lower level, the Data Layer sub-component comprises all the tools deeply linked to input data and allows for the flow of information to and from the platform. The component is also responsible for the interaction with external data sources and the proper integration with the rest of the ingested information. The Data Layer component is at the core of this deliverable and will be further expanded in the next sections with a more in-depth description about each individual tool and data source involved. As for the AI-Based Advisory Tool, an extensive presentation concerning the methodology behind it can be found in D2.2, while the technical details will be presented as part of deliverables D4.9 and D4.10.

One level of abstraction upper, the Service Layer comprises a set of different tools that relies on the available data to produce new knowledge and exposes a set of services that will be exploited by different actors in the upper layer. This sub-component is constituted by tools coming from different backgrounds (technological, ecological, ethnographical, and operational), integrated at different levels in order to accommodate the platform needs. For a more exhaustive description of each component, as well as the detailed integration schema, please refer to D4.1.

Lastly, the upper level is constituted by the three main platform modules: the Disassemblability Module (DIS), the Recyclability Module (REC) and the Eco Design Module (ECO). Each one of them exploits different functionalities offered by the underlying components to generate value to the different actors involved in each module. More specifically, the Disassemblability Module will focus on the dismantling process of EoL vehicles both from a methodological and industrial point of view, the Recyclability Module will be centred around recycling procedures for materials extracted from EoL components, while the Eco Design Module will support decision making processes for BoL actors responsible for components and car parts manufacturing.

2.2 Components description

As already mentioned in the previous section, the main components relevant in the scope of D4.3 are located into the Data Layer sub-component; those are the Data Lake and the Data Importer. It is important to notice that, even if the AI-based Advisory tool is already foreseen as a part of the Data Layer, its functionalities will be expanded, finalized and described within D4.9 and D4.10, leveraging on the advisory methodology defined in D2.2.

Here below are described the main functionalities of the Data Lake and Data Importer components:

- Data Lake: this component acts as a centralized knowledge base in which a wide variety of data sources are grouped together. The resulting data is a combination of the output of different tools constituting the TREASURE Platform, knowledge provided by involved technical partners and additional information coming from industry-standard external data sources. The complete set of information contained in the TREASURE Data Lake are the following:
  - Economical, technological and circularity information, coming from external data sources (IDIS database, MISS database, Insurances database).
  - Augmented reality procedures, execution logs and user feedback, coming from the WEAVR Platform.
  - Semantic information, ethnographic data and user sentiment analysis, coming from the SSNA Platform.
  - Information about car parts, components, materials and dismantling instructions, coming from BoL actors (cars/components manufacturers) and EoL actors (car parts/components disassemblers and dismantlers).
  - Environmental KPIs, materials life cycle information and recycling best-practices, coming from the Sustainability Tool (leveraging external data sources such as the Ecoinvent database) and EoL actors (car parts/components shredders and components/materials recyclers).
  - Recycling KPIs and recycling/disassembly routes coming from the Recycling Simulation Tool.
- Data Importer: this component is responsible for realizing the data integration with external sources. The TREASURE Platform will indeed rely on information coming from well-known CE data providers in order to enrich the amount of data at disposal of the various tools. This will allow to have a more complete understanding about the impact

of circular economy and circularity best practices in the automotive industry. To achieve this result, the outer data provided are ingested in the TREASURE Data Lake and converted in a suitable standardized format. This integration with external data sources is the result of a joint effort between TXT and technical partners who retain access to those external data sources; in particular the integration of IDIS was realized with the contribution of ILSSA, the integration of Insurances DB was made possible thanks to POLIMI and integration of MISS thanks to SEAT via ILSSA.

2.3 Actors involved

This chapter provides information about the actors involved in the Data Lake from different point of views. More specifically, all the actors whose actions are related with the flow of data through the Data Lake component are considered, including the final utilizers of such data, the users who brings knowledge into the platform and the research and innovation partners that study the available information from an academical perspective. Below it is depicted the list of involved actors, along with their role with respect to the Data Lake:

- Dismantlers, shredders and physical operators (ILLSA, POLLINI): leverage knowledge from the platform in order to carry out their designed tasks. This is achieved by exploiting other parts of the TREASURE Platform, such as the WEAVR, and the cobot.
- Car makers, components manufacturers and parts suppliers (SEAT via UNIZAR, EUROLCDS, POLLINI): provide knowledge to the platform by uploading car parts/components information into the centralized shared space.
- External data sources providers (ILSSA, POLIMI, POLLINI, SEAT): contribute to enlarge the content of the Data Lake by providing additional information about external data sources (IDIS, MISS, Insurances DB).
- Recycling operators (ILLSA, UNIVAQ): leverage the information available in the Data Lake to perform best recycling procedures, according to relevant KPIs. These activities are performed with the usage of other tools offered by the TREASURE Platform, such as the WAVR, Recycling Simulation Tool, and Sustainability Tool.
- Supportive partners (TXT, POLIMI, SUPSI, MARAS, SEAT, UNIZAR, EDGE): provide different kinds of data to be integrated in the Data Lake, contributing to enlarge the knowledge base at disposal of other partners/tools:
    - TXT: provides information about dismantling/recycling procedures, execution logs, training sessions and operators' feedback.
    - TXT, POLIMI: provide joint information thanks to the integration of the COBOT with the WEAVR Platform in supporting dismantling activities.
    - SUPSI: provides sustainability KPIs and impacts of processes and materials as a result of the output from the Sustainability Tool.
    - MARAS, SEAT via UNIZAR: provides the optimal recycling processes, recycling routes and materials information, thanks to the analysis performed by the Recycling Simulation Tool and knowledge gathered from external data sources.
    - EDGE: provides information about the semantic social impact of CE practices, as well as user sentiment analysis on circularity topics. This is achieved using the SSNA Tool and its data output.
- Research partners (TNO, UNIZAR, UNIVAQ): access information about in-mold electronics shared by other partners and elaborated through the TREASURE Platform,

in order to develop novel prototyping processes and further discuss IMSE (In-Mold Structural Electronics) adoption from a methodological point of view.

3 Data Lake

This chapter focuses on the description of the TREASURE Data Lake platform component. The section starts summarising the overall information about the software released (description, overall data, functionalities and architecture), followed by technical information about architectural stack, technical manual for installation and licensing (including third parties' components). Finally, the conclusions and future steps close the chapter.

3.1 Component Description

3.1.1 Overall Data

Item	Value
Component Name	Data Lake
Software version	V 1.0.0
Reference workpackage	WP4
Responsible Partner	TXT
Contact person	Mattia Calabresi, mattia.calabresi@txtgroup.com
Source control	No publicly available source control for the Data Lake component
Short Description	The TREASURE Data Lake component offers a centralized platform to access knowledge, provided by different actors and gathered from multiple data sources, spanning a wide variety of data types

3.1.2 Purpose of the component

The main objective of the Data Lake component is to stand as the centralized knowledge base inside the TREASURE Platform. In particular, this component aims at providing data storage for all the information that are considered relevant in the scope of the project, hosting data coming from different tools, actors and data sources (both internal and external with respect to the project's scope). Furthermore, the Data Lake is responsible for making this knowledge accessible to all components of the platform and, through them, also to the end users.

3.1.3 Summary of Functionalities

This section focuses on describing the main functionalities offered by the Data Lake component, with a high-level view on the principal characteristics for each one of them. The main interfaces offered by the component are:

- Provide data access to other platform components by offering a standardized, query-based format to retrieve information.
- Provide data access to end users either directly or by means of dedicated platform tools. An example of the latter is the ability of EoL actors to retrieve suggestions, best recycling routes and CE KPIs through the Circularity Web Platform and its dedicated dashboards.
- Provide storage of data coming from other tools as output of their processes. This, for instance, is realized by establishing a communication protocol between a producer component (e.g., the WEAVR Platform, Recycling Simulation Tool, …) and a dedicated storage section of the Data Lake.
- Provide storage of data coming from authorized users that decides to share relevant information with the rest of the platform, contributing to enhance the overall

knowledge at disposal of TREASURE. An example of this can be found in the contribution of BoL actors (car/component manufacturers) sharing recycling/manufacturing data with the Data Lake component.

- Regulate access of users depending on their role, organization and sensitivity level of information requested. This will be realized by the integration of the platform with a RBAC (Role-Based Access Control) system, able to provide isolation of information and grant access to sensitive data only to the authorized users.

3.2 Technical Information

3.2.1 Internal Architecture

From an architectural point of view, the Data Lake component is realized leveraging state of the art storage technologies offered by the AWS platform. The technical stack is split into different components, each one providing fast, reliable and secure storage solutions depending on the type of data to be managed. In particular, for non-relational data types, the preferred storage solution is through MongoBD in JSON format. To do so, the AWS Document DB component is the preferred choice. DocumentDB is a database service that is purpose-built for JSON data management at scale, fully managed and integrated with AWS, and enterprise-ready with high durability. Concerning structured data instead, the relational database of choice is PosgreSQL for the storage of data in tabular format. For this use case, the selected AWS technology is Amazon RDS. RDS is a managed relational database service that provides a wide variety of database engines to choose from, including Amazon Aurora, MySQL, MariaDB, Oracle, Microsoft SQL Server, and PostgreSQL.

3.2.2 Technological stack

Item	Value
Nature	Distributed data store
Programming Language	MySQL, Postgres, MongoDB
Development Tools	MySQL Workbench, pgAdmin, AWS Cloud Console
Additional Libraries	No additional libraries required
Databases	MongoDB and PostgreSQL depending on the data to be stored. Several data sources are involved in the Data Lake. For an exhaustive list of them, please refer to chapter 3.2.3

3.2.3 Data Sources

This chapter focuses on presenting the different data sources that contribute to populate the TREASURE Data Lake. In the scope of this project, several kinds of data sources have been taken into account, ranging from output produced by other platform tools, direct user input and information from external sources.

3.2.3.1 IDIS DATABASE

Description

IDIS is the advanced and comprehensive information system for pre-treatment and dismantling information for End-of-Life Vehicles (ELV). It contains safe handling information with focus on airbag deployment instructions, handling and treatment of high voltage batteries as well as gas vehicles. Additionally, it provides user friendly navigation to an extensive database with practical information on pre-treatment, dismantling of potentially recyclable parts and other elements mentioned in ELV regulations (e.g., mercury, lead, cadmium and chromium VI). The database contains relevant data to be processed by TRASURE platform mainly with respect to disassemblability module.

Data structure

Describe the format in which data are stored (e.g., plain text, JSON, tabular format, ...) and the most important features about the data source.

Sample Datatypes

When possible, provide a sample of the stored data (copy-paste text, screenshots, links, ...).

3.2.3.2 MISS DATABASE

Description

MISS (Material Information Sheet System) is an internal software belonging to the Volkswagen (VW) Group, which obtains the information from IMDS (International Material Data System). The IMDS was created due to the ELV (End of Vehicle) Directive, for accomplishing with the legislation requirements. All automotive car part manufacturers are obligated to provide the information regarding the material composition and weight of the car parts under its responsibility. MISS system contains the material composition and weight of the car parts used in VW Group cars, displaying the information in a tree structure from the main car part up to the elemental level.

Data structure

Since a direct connection between miss and treasure is not possible, the information provided by the external database will be converted in a data format that the platform can process.

3.2.3.3 Insurances DATABASE

Description

The Insurances Database is a private data store, used by insurance companies in order to obtain information about the repairing process of damaged vehicles. The platform allows operators to identify the vehicle using the chassis manufacturer number providing access to repair times, cost of replacement parts, cost of labor and time to repair.

Data structure

Data from Insurances DB are retrieved in two ways:

1. In PDF format leveraging Audatex: the largest database software in the world for the identification and repair of vehicles in the bodywork and mechatronics field. The software gives information about disassembly time, economic value of the car part, exploited images of the location where the car part is installed and reparation times.
2. Via web APIs with a pay-per-use subscription model, gives access to more detailed information in standardized JSON format. This channel is still under discussion, and it is not yet available for the time being.

3.2.3.4 WEAVR Platform

Description

In the scope of the TREASURE Project, the WEAVR Platform aims at improving the activities performed in the DIS and REC modules, providing physical operators (shredders and dismantlers for the former, recyclers for the latter) with the set of virtual reality tools designed to manage, execute and improve AR/VR procedures for the disassembly and recycling of car parts/components.

During this entire process, a wide variety of data is produced, analyzed and stored in order to allow the platform to perform the mentioned tasks. In particular, the data sources the WEAVR Platform relies on is composed by four databases, each one exploiting a core functionality of the infrastructure:

- IdentityDb: contains all the information related to the users operating inside the WEAVR Platform, their roles, related organizations, membership groups and access credentials.
- ContentDb: contains all the metadata related to the procedures and their composing steps, grouped by supported platform (Windows, Mac, ...) and execution device (Android, Oculus, HoloLens, ...)
- AnalyticDb: includes all the information recorded during the execution of each step, for every procedure and user. It contains records about the session duration, completed and failed steps, etc...
- LTIConnector: this DB contains metadata that allow the WEAVR Platform to be integrated with external wikies and third parties' visualization tools, such as the Moodle management system.

More details about the information contained in each individual data source will be provided in the next section.

Concerning the architecture and deployment model, each data source is individually containerized and offered through the AWS cloud platform. In particular, the selected hosting solution uses Amazon ECS and ECR services to automate the management of containerized applications. ECS (Elastic Container Service) is a container orchestrator used to quickly deploy clustered applications with minimal setup required, while ECR (Elastic Container Registry) offers a secure and reliable data source to upload container images to be served by ECS. These services offer critical functionalities such as auto-scale, load balancing and automatic fail recovery with minimal setup and high reliability.

Data Structure

The data structure used by the WEAVR Platform comprises many different kinds of information, concerning all aspects of the users' activities inside the platform. For this reason, a combination of relational and non-relational data store has been put in place, allowing to efficiently store procedures executable files as well as metadata, logs and user profiles. In particular, the data collected for WEAVR are listed below by topic:

- Users' information (name, email, password, department, company, ...) including their role into the WEAVR Platform that regulates permissions for specified resources following an RBAC access model.
- Groups: collections of users belonging to the same department. It is used to allow administrators to perform collective operations on all the members simultaneously. In particular, for each group, it keeps track of the list of participants.
- Companies: collections of users belonging to the same organization, similar functionality offered by groups, at a higher level. Also, similar to the previous entry, the list of participants for each company is stored.
- Procedures: training, simulation procedures and operations executed by users. Those are represented in the form of binary files which are platform-dependent and device-specific. For this reason, an individual version of each procedure exists, for each device it has to be run on. Procedures are hosted in a non-relational data store (AWS S3 bucket) and referenced in the metadata DB. Also, procedures are accompanied by their relative metadata such as supported execution platforms, interaction modalities, activity type, number of executions and assigned users/groups/companies.

- Logs (also called analytics): metadata collected automatically during the procedure execution; those include number of repetitions, time taken by each user to complete each step, steps failed and succeeded, as well as notes taken by the users during the simulation.
- Licenses: metadata related to the WEAVR platform utilization privileges, including the users license ID and expiration date.

Sample Datatypes

Here below are reported some examples of non-sensitive data produced and stored by the WEAVR Platform.

	Id [PK] uuid	AuthUserId uuid	Email character varying (100)	FirstName character varying (100)	LastName character varying (100)	Department character varying (100)	LastLoginAt timestamp without time zone	StatusId uuid	CompanyId uuid	CreatedAt timestamp without time zone	CreatedBy uuid
1	fd01235a-...	fd01235a-ae6...	lorenzo.mureddu@txtgroup....	Lorenzo	Mureddu		[null]	ff4334aa-...	0d5c4a5e-00...	2022-02-17 10:49:47.414878	e4828aab-0...
2	60e692ad-...	60e692ad-48...	lorenzo.mureddu@txtgroup....	Lorenzo	Mureddu		2022-02-17 14:43:39.762211	9b763d4d...	0d5c4a5e-00...	2022-02-17 13:50:14.266027	e4828aab-0...
3	e4828aab-...	e4828aab-00...	superadmin@weavr.com	SuperAdmin	User	WEAVR	2022-02-17 14:55:50.671033	9b763d4d...	0d5c4a5e-00...	2022-02-11 08:43:19.07386	00000000-0...
4	b44d4cc0-...	b44d4cc0-27...	mattia.calabresi@txtgroup....	Mattia	Calabresi		2022-04-12 14:43:36.138685	9b763d4d...	0d5c4a5e-00...	2022-02-17 11:42:26.752677	e4828aab-0...

Figure 2 Sample of users' data

	Id [PK] uuid	UnityId uuid	Name character varying (100)	Description character varying (1000)	Configuration character varying (10)	Status integer	Ownership integer	SceneId uuid	CompanyId uuid	CreatedAt timestamp without time zone	CreatedBy uuid	UpdatedAt timestamp without time zone	Updat uuid
1	164799cc-...	0000000...	737 Landing	Training for emergency landi...	VT	0	1	c5f43b22-...	71792301-b8...	2022-02-11 08:43:12.866272	00000000-0...	2022-02-11 08:43:12.866272	00000
2	b0d6682c-...	0000000...	737 Departure check	Training for 737 departure du...	OPS	0	1	c5f43b22-...	71792301-b8...	2022-02-11 08:43:13.072979	00000000-0...	2022-02-11 08:43:13.072979	00000
3	d6bfc658-...	0000000...	737 Door Closing	Training for shipboard person...	VT	0	1	678afbc8-...	71792301-b8...	2022-02-11 08:43:13.073662	00000000-0...	2022-02-11 08:43:13.073662	00000
4	4bdfe257-...	0000000...	737 Door Closing	Training for shipboard person...	OPS	0	1	678afbc8-...	71792301-b8...	2022-02-11 08:43:13.074101	00000000-0...	2022-02-11 08:43:13.074101	00000
5	94e659a7-...	0000000...	A320 Departure	Training for A320 departure, ...	OPS	0	1	74432281...	71792301-b8...	2022-02-11 08:43:13.074476	00000000-0...	2022-02-11 08:43:13.074476	00000
6	78240b21-...	0000000...	A320 Departure	Training for A320 departure, ...	VT	0	1	74432281...	71792301-b8...	2022-02-11 08:43:13.074844	00000000-0...	2022-02-11 08:43:13.074844	00000
7	847d4b06-...	23352af...	DEMO Interaction Scene VT	The example procedure with ...	VT	0	0	1b5d0376...	0d5c4a5e-00...	2022-02-17 11:43:40.287971	e4828aab-0...	2022-02-17 11:43:40.287971	e4828
8	4e87b0b3-...	097db28...	DEMO Interaction Scene OPS	Demo OPS procedure	OpS	0	0	1b5d0376...	0d5c4a5e-00...	2022-02-17 11:45:38.749201	e4828aab-0...	2022-02-17 11:45:38.749201	e4828

Figure 3 Sample of procedures data

	Id [PK] uuid	UnityId uuid	Name character varying (100)	CompanyId uuid	CreatedAt timestamp without time zone	CreatedBy uuid	UpdatedAt timestamp without time zone	UpdatedBy uuid	DeletedAt timestamp without time zone	DeletedBy uuid	ProcedureTokenId uuid	UploadStatus integer
1	c5f43b22-...	0000000...	Cockpit 737	71792301-b8...	2022-02-11 08:43:12.674452	00000000-0...	2022-02-11 08:43:12.674452	00000000-00...	[null]	[null]	00000000-0000-0000-...	
2	678afbc8-...	0000000...	Inside 737	71792301-b8...	2022-02-11 08:43:12.714394	00000000-0...	2022-02-11 08:43:12.714394	00000000-00...	[null]	[null]	00000000-0000-0000-...	
3	74432281-...	0000000...	Cockpit A320	71792301-b8...	2022-02-11 08:43:12.714599	00000000-0...	2022-02-11 08:43:12.714599	00000000-00...	[null]	[null]	00000000-0000-0000-...	
4	1b5d0376-...	e33938c...	Assets/WEAVR/Demo/Sce...	0d5c4a5e-00...	2022-02-17 11:43:39.727496	e4828aab-0...	2022-02-17 11:45:38.593469	e4828aab-00...	[null]	[null]	88e88fa6-19d5-42d1-...	

Figure 4 Sample of scenes data

3.2.3.5 SSNA Tool

Description

The main content of the data source is the open conversation of the car consumers on the circular economy in the automotive industry, collected and fostered by Edgeryders ethnographers and community management.

The tools used are the following:

- Edgeryders online platform
- its embedded tool for ethnographic coding called OpenEthnographer
- the visualisation tool GraphRyder and its interactive dashboard.

The data store is publicly accessible.

Data structure

The data will be exported in a widely used format (JSON) and stored for long-term safekeeping (e.g., Zenodo repository).

Sample Datatypes

The visualisation of the first batch of data is available through the GraphRyder interactive dashboard.

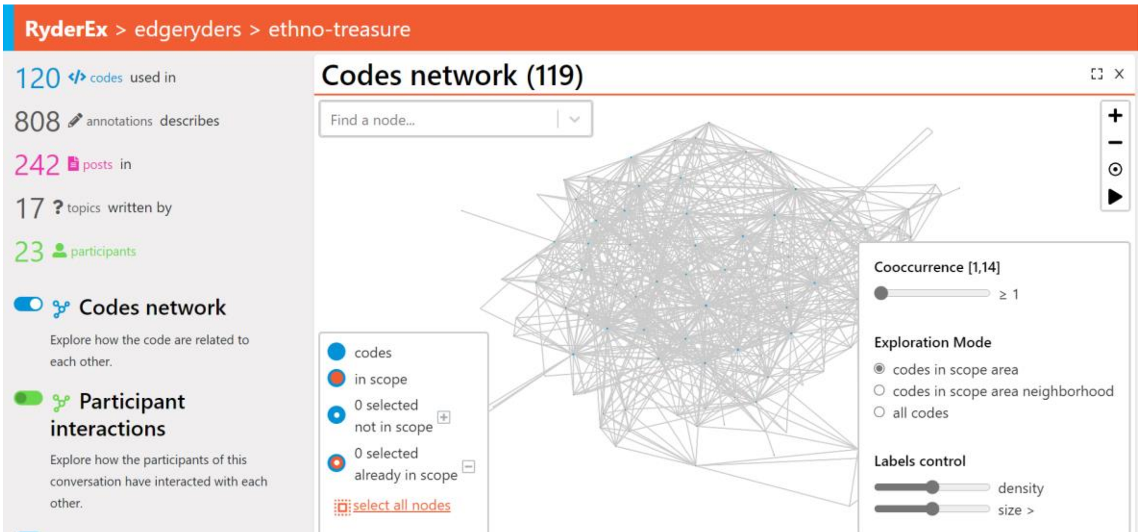

Figure 5 GraphRyder interactive dashboard

A first analysis of the co-occurrence network has been performed and is available here.

The GraphRyder technical manual is available here.

3.2.3.6 Sustainability Tool

Description

The objective of the Sustainability Tool, in the scope of TREASURE, is to allow the calculation of the environmental, social, economic, and circular indicators selected taking into account relevant existing standards and the Life Cycle Sustainability & Circularity Assessment (LCS&CA) methodology developed in TREASURE T2.1. The indicators will address not only the specific sustainability & circularity area, but they will be categorized based on the specific life cycle phase they are evaluating, namely use case dedicated indicators will be provided, with major focus on EoL phases. Those indicators, calculated by the Sustainability Tool, will feed the knowledge base on which the AI-based Advisory Tool is based on, which will be defined upon the sustainability advisory methodology developed in TREASURE T2.2. In this regard, this section is meant to describe the first version of the data model on which the Sustainability Tool functionalities are based on. In particular, the data model described below is mainly focused on the LCA aspects and allows formalizing industrial processes at the environmental sustainability point of view.

Data structure

The data model of the Sustainability Tool is based on the concepts widely described within the deliverable D4.1. In order to guarantee a complete comprehension to the reader, the main concepts are hereafter summarized:

- ***Process template*** is a simplified description of the manufacturing operations from a sustainability perspective. In particular, process templates are meant to formalise the LCI description of the processes where inputs and outputs flows are identified and quantified.
- ***Input flow***: from the input side, LCI considers resources coming from the eco-sphere (e.g., raw material, water) or from another techno-sphere (e.g., electricity mix, ancillary

material such as lubricating oil, etc.) and energy of various types. An input flow can be considered as an "*elementary flow*" when the exchange occurs with the natural, social, or economic environment (e.g., un-processed inputs from nature, working hours under specified conditions, etc.).

- ***Output flow***: from the output side, LCI inventories consider emissions (directed to the different environmental compartments), waste, products (i.e., the results of the system in analysis) and co-products. As for the input flows, the output flow can also be considered an "elementary flow" when the exchange takes place with the natural, social or economic environment (e.g., emissions to air, water and soil, physical impacts, etc.).

Concerning the LCA model, LCI data represents the input variables of the process. LCI data can be retrieved directly from the production line and can be collected manually or through IoT devices. These kinds of LCI information are called foreground or primary data. Besides, LCI can also be obtained from databases (i.e., Ecoinvent), or alternatively literature or statistical data. These kinds of LCI information are called secondary or background data.
Figure 6 below shows the main entities belonging to the first version of the data model behind the Sustainability Tool, which is mainly focused on the LCA aspects. This data model has been organized in two main groups of entities:

- *LCA Process Characterization*: this group contains all entities needed to formalize the LCA-oriented characterization of processes.
- *Sustainability Assessment*: this group contains all entities needed to formalize the sustainability assessment, performed by means of a specific LCA methodology, related to a specific process.

Hereafter, the main entities of the data model are briefly explained for both the groups.

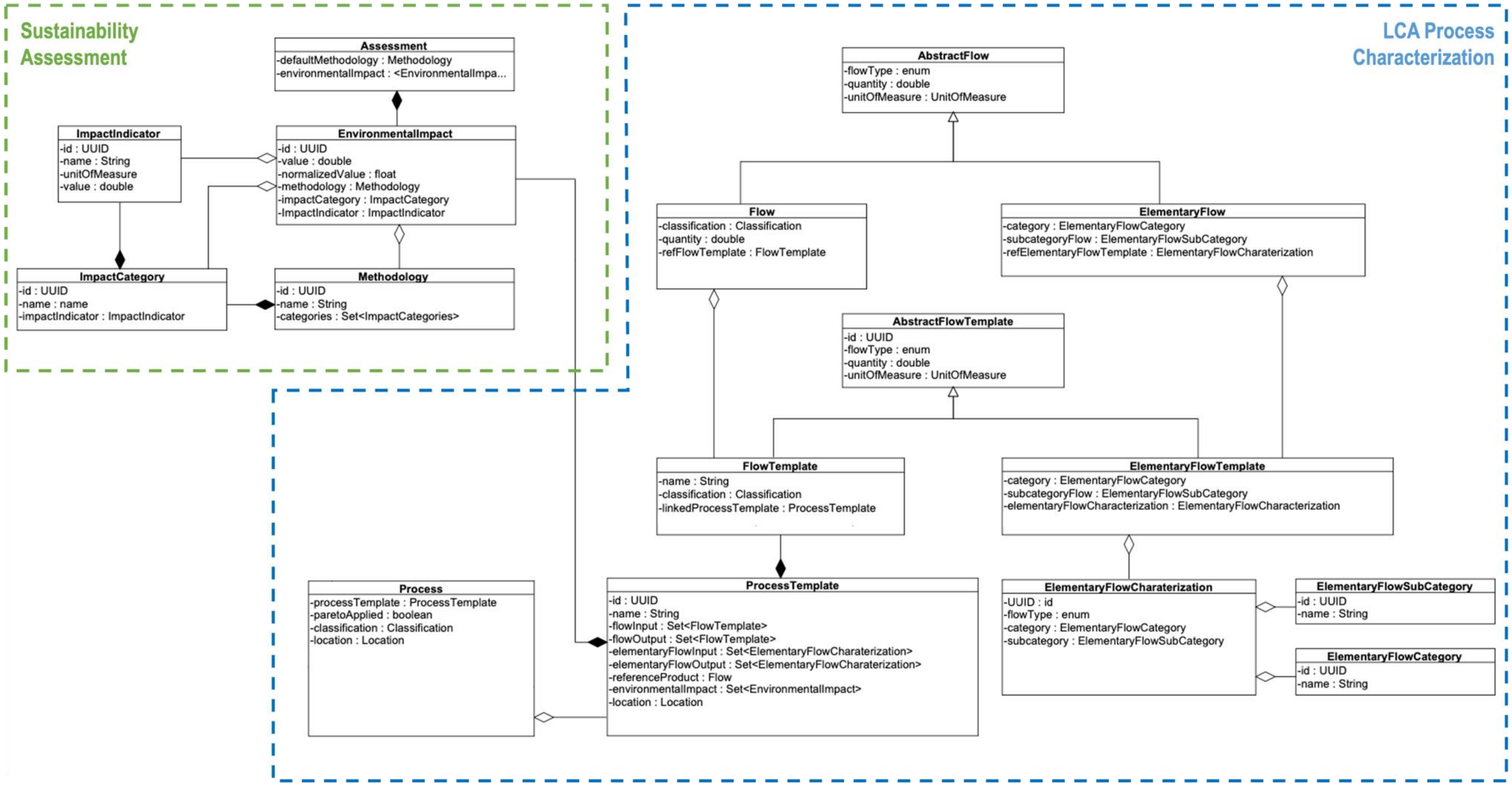

Figure 6 Sustainability Tool Data Model (V1)

LCA Process Characterization

Entity	Description
Process	This entity represents an instance of a process and represents a customization of a specific process template.
Process template	This entity is meant to describe a manufacturing operation. It contains a set of attributes needed to describe a process from a sustainability perspective: - *name*: name of the process template (i.e., "injection moulding", "aluminium drilling", etc.) - *flowInput*: set of input flows belonging to the process template. - *flowOutput*: set of output flows belonging to the process template. - *elementaryFlowInput*: set of elementary input flows belonging to the process template. - *elementaryFlowOutput*: set of elementary output flows belonging to the process template. - *environmentalImpact*: set of environmental impacts calculated by means of a selected LCA methodology.
AbstractFlowTemplate	This entity represents a generalization of the concept of flow template and contains all data common to all subtypes of flow templates (such as type, quantity and unit of measure).
FlowTemplate	This entity is a subtype of the *AbstractFlowTemplate* entity related to a specific process template.
ElementaryFlowTemplate	This entity is a subtype of *AbstractFlowTemplate* and it is configured as an elementary flow.
AbstractFlow	This entity represents a generalization of the concept of flow and contains all data common to all subtypes of flows (such as type, quantity and unit of measure).
Flow	This entity is an instance of a flow and represents a customization of a specific *FlowTemplate*.
ElementaryFlow	This entity is an instance of an elementary flow and represents a customization of a specific *ElementaryFlowTemplate*.

Sustainability Assessment

Entity	Description
Assessment	This entity is meant to describe a sustainability assessment and contains the set of calculated environmental impacts and the methodology used for the calculus.
EnvironmentalImpact	This entity is an instance of a single environmental impact and contains a specific impact indicator calculated by means of a specific LCA methodology.
ImpactIndicator	This entity contains the impact value calculated by the sustainability tool referred to the customized process. It is identified by name and contains value and unit of measure.

Methodology	This entity is meant to describe the LCA methodology used for performing a sustainability assessment.

The Sustainability Tool manages the exchanged data, both between the front-end and back-end and the back end with the TREASURE platform, by means of JSON objects (one for each involved production process) meant to describe the process at sustainability point of view. In particular, such JSON file describes the list of involved processes templates and the related LCI (e.g., “electricity”, “lubricating oil”, “quantity of material”, etc.) that actually affect most of the process’ sustainability impacts.

Sample Datatypes

The JSON file depicted below in Figure 7 is meant to describe the injection moulding process according to the data model previously presented.


```
{
   "processes": [
    {
      "id": "f66508d8-92bb-4605-8701-55026d9cfb6c",
      "processTemplateId": "a25277164a5837f2bdb...",
      "referenceProductId": "68d6dfcf-2089-4586-9bbf-ad75591105cf",
      "processBaseName": "injection molding",
      "referenceProductName": "injection molding",
      "locationAreaName": "RER",
      "paretoApplied": true,

      "flows": [
        ...
      ],

      "elementaryFlows": [
        ...
      ],

      "mixes": [
        ...
      ]
    }
  ]
}
```


Figure 7 Sample: injection molding process characterization

```
"flows": [
  {
    "id": "e903ad14-c719-4aed-b9f6-7d8e97bed81f",
    "name": "heat, district or industrial, natural gas",
    "flowType": "FROM_TECHNOSPHERE",
    "flowDirection": "INPUT",
    "originalQuantity": 4.21,
    "quantity": 4.21,
    "quantityFactor": 1,
    "unitOfMeasureName": "MJ",
    "impactsContribute": 0.13129,
    "flowTemplateId": "b014b4b6f17d01e24cd26...",
    "geographyShortName": "RER"
  },
  {
    "id": "85fd7b4b-c324-4cef-a0bb-58e9f6edacef",
    "name": "injection molding",
    "flowType": "REFERENCE_PRODUCT",
    "flowDirection": "OUTPUT",
    "originalQuantity": 1,
    "quantity": 1,
    "quantityFactor": 1,
    "unitOfMeasureName": "kg",
    "impactsContribute": 0,
    "flowTemplateId": "95213cd5c98f40906f2dfb88...",
    "geographyShortName": "-",
  },
  ...
 ],
"elementaryFlows": [
  {
    "id": "ce869f95-9607-47e8-a47b-7ff0b8036bf7",
    "name": "Water",
    "flowType": "TO_ENVIRONMENT",
    "flowDirection": "OUTPUT",
    "originalQuantity": 0.0042625,
    "quantity": 0.0042625,
    "quantityFactor": 1,
    "unitOfMeasureName": "m3",
    "impactsContribute": 0,
    "paretoIncluded": false,
    "elementaryFlowTemplateId": "a175e14a646073907d1cb03..."
  },
  ...
 ]
```


Figure 8 Sample: input and output flows related to the injection molding process

```
"mixes": [
  {
    "id": "66dd9b6c-8b51-41ec-9554-28555b60ccd3",
    "name": "Electricity mix",
    "mixBaseName": "Electricity mix",
    "templateType": "ElectricityMixTemplate",
    "originalQuantity": 1.48,
    "quantity": 1.48,
    "quantityFactor": 1,
    "unitOfMeasureName": "kWh",
    "impactsContribute": 0.69229
  }
]
```

Figure 9 Sample: electricity mix related to the injection molding process

```
"assessment": {
      "methodology":
      {
    "methodologyId": "679a3f0b-79df-4527-af9e-9ff901114715",
    "methodologyName": "IMPACT 2002+ (Endpoint)"
  }
  "impacts": [
        {
          "methodology": "IMPACT 2002+ (Endpoint)",
          "category": "climate change",
          "indicator": "total",
          "value": 0.000090586455,
          "unit of measure": "points",
          ...
        },
        {
          "methodology": "IMPACT 2002+ (Endpoint)",
          "category": "ecosystem quality",
          "indicator": "terrestrial ecotoxicity",
          "value": 0.000009401759,
          "unit of measure": "points",
          ...
        },
      ...
  ]
}
```

Figure 10 Sample: sustainability assessment of the injection molding process

3.2.3.7 Recycling Simulation Tool

Description

Within the TREASURE project, the Recycling Simulation Tool as developed by MARAS will be applied to calculate recycling/recovery rates of the disassembled components as well of the IME's. Hence, it allows for quantification of recovery rates of the entire disassembled part, as well of its composing elements, materials and compounds (implying material defined in its stoichiometric chemical composition, i.e., aluminum as Al, Al2O3, full composition of organic molecules of C, H, O, N, Br, Cl, metals atoms etc.). At the same time, most optimal recycling system infrastructures will be defined (i.e., disassembly depth and most optimal combination of recycling processes).

The data source as input for the recycling simulations is the product and part compositional data provided in the format of the MISS, which is being delivered as pdf by the automotive OEMs or as xlsx file for the IMEs. In order to derive the required input detail from this type of data, the data is being processed by MARAS in order to set up xlsx files for the different parts, containing a structured (linked to product tree) list of all materials and corresponding compound stoichiometric formulas for each product part and subpart and its spatial distribution in the product/part and all corresponding individual weights. This is the complete composition of the product, thus all compounds, functional materials, alloys, plastics etc. and their spatial position on the modules and their masses. This means aluminum in Al, an alloy of aluminum, Al2O3 as an oxidized/anodized layer on the aluminum, or a filler etc. captured in xlsx files. This data contains the feed composition as input to the recycling simulation model, which requires a full description of the compounds, which must add up to 100% in weight.

The data processing is an extremely labor intensive, but needs to be done, as the MISS data files do not contain the required data detail for recycling assessment. This should be automated in the platform in order to allow for a smooth data conversion between OEM data and recycling rate predictions.

The flowsheets are contained in the model structure and code and mean a logical sequence of reactors that convert the input into among others high quality materials, compounds, alloys, metals, building materials, energy as well as residues and intermediates that can be ponded ore used in further processes. These flowsheets are industrially realistic and economically viable for different processing routes. All the flows of materials, solution, mixture, phases, gases, dust (among others) are quantified within the model structure and code in terms of enthalpy and entropy (this kWh/h) values in addition to the mass flows (both total and compound flows) in kg/h or tons/h.

The output data of the models are delivered in xlsx format when referring to product, part as well as individual material, element and compound recycling rates.

The flows and results of the recycling simulations are expressed in xlsx data format in terms of enthalpy and entropy (this kWh/h) values in addition to the mass flows (both total and compound flows) in kg/h or tons/h.

The output in terms of most optimal processing structures is provided in written text format and/or flowsheet graphics.

Data structure

The input data is delivered either as pdf or xlsx data file by the other parties in the project. This input is converted by MARAS to xlsx format as input to the simulation models.

The output is generated in the model and converted to xlsx data format. Flowsheet graphics and written instructions provide the data output in terms of recycling instructions.

Sample Datatypes

Below some examples of the input data (MISS), converted input data (xlsx), model data and output data (xlsx) are given.

Teilecharakterisierung

Baum-Ebene	Typ	Name	Teilnummer/ Sachnummer/ Werkstoffnummer/ CAS-Nummer	Menge	Gewicht [g]	Mengenanteil [%]	Mengenanteil [%] (von - bis)	VDA-Kategorie/ Reinstoff-Eigenschaften	Polymer-Kennzeichnung/ Rezyklat (Prod.-Abfall/Altmetall)/ Anwendung	SCIP-Informationen
1	■	SEAT PQ35 Radio Low RCD 310 UP2	5P0 035 186 B	1	1973,00					
|—2	■	Case/Gehäuse	-	1	1098,00					
|—3	■	Metall	-	1	880,00					
|—4	○	Screw-Material (Torx)	-		7,91					
|—5	●●	Material for Fasteners Property Class <=12.9 (Flat Bill)	-			99,656		1.1.1	Rezyklat enthalten: Nein	
|—6	▲	Kohlenstoff	7440-44-0			0,275	0.0 - 0.55			
|—6	▲	Phosphor	7723-14-0			0,0125	0.0 - 0.025			
|—6	▲	Schwefel	7704-34-9			0,0125	0.0 - 0.025			
|—6	●●	PCB-epoxy for PCB laminate, High Component Load	-			21,649573	19.0 - 24.0	5.4.3	Rezyklat enthalten: Nein	
|—7	▲	Epoxidharz	-			86,70				
|—7	▲	Acrylate	-			0,25	0.0 - 0.5			
|—7	▲	Bariumsulfat	7727-43-7			0,10	0.0 - 0.2			
|—7	▲	TBBA	79-94-7			7,50	5.0 - 10.0	deklarationspflichtig		
|—7	▲	Titandioxid	13463-67-7			0,50	0.0 - 1.0			
|—7	▲	Talk	14807-96-6			0,20	0.0 - 0.4	deklarationspflichtig		
|—7	▲	Sonstiges, nicht zu deklarieren	system			2,00	1.0 - 3.0			
|—7	▲	Pigmentanteil, nicht zu deklarieren	system			1,75	0.5 - 3.0			
|—7	▲	Siliciumdioxid, glasartig	60676-86-0			1,00	0.0 - 2.0			
|—6	●●	PCB-epoxy for components, High Component Load	-			6,589744	5.0 - 8.0	5.4.3	Rezyklat enthalten: Nein	
|—7	▲	Bismut	7440-69-9			0,05	0.0 - 0.1			

Figure 11 Sample of MISS input data as obtained from OEM (selected cut out in view of confidentiality) (part of cas numbers lacking, no stochiometric composition and no mass distribution of entire part, only distribution in % of mass per subpart) (supplied in pdf

Input streams	Value	Units	Flow Rates			Thermal E Flow	Total H Flow
			kg/h	Nm³/h	kmol/h	kW	kW
Total Gas Flow	0.00	Nm3/h					
Total Condensed Flow	117.74	t/h	117 738.87	#DIV/0!	972.43	- 1.14	-286 452.83
Treasure Module 1	20.00	t/h	Flow Rates			Thermal E Flow	Total H Flow
Temperature	25.00	°C	kg/h	Nm³/h	kmol/h	kW	kW
Pressure	1.00	bar					
Total	100.00	wt-%	20 000.00	#DIV/0!	205.09	0.61	-36 681.02
Al2O3	23.69		4 737.61	1.19	46.46	0.00	-21 627.95
C11H30O3Si4	0.00		0.26	#DIV/0!	0.00	0.00	0.00
C10H8O4	0.07		13.65	#DIV/0!	0.07	0.00	- 14.82
C12H12O(1ENg)	0.87		173.64	22.90	1.01	0.00	- 14.76
C12H22N2O2	0.02		4.32	#DIV/0!	0.02	0.00	- 2.01
C22H10N2O5	0.00		0.49	#DIV/0!	0.00	0.00	- 0.25
C3F6O(HFAg)	0.00		0.08	0.01	0.00	0.00	- 0.19
C3H4O2	0.00		0.65	#DIV/0!	0.01	0.00	- 1.10
C3H6	10.09		2 017.69	#DIV/0!	47.95	0.00	-1 175.95
C5H8O2	14.44		2 887.57	#DIV/0!	28.84	0.00	-3 950.97
C12H11N(4AB)	22.30		4 460.52	#DIV/0!	26.36	0.00	592.94
C7H5O(g)	0.01		1.37	0.30	0.01	0.00	0.42
TiO2	0.00		0.09	0.00	0.00	0.00	- 0.31
C4H6O4(SUC)	0.02		4.20	0.00	0.04	0.00	- 9.28
Ag	0.06		11.27	0.00	0.10	0.00	0.00
Al	0.15		29.52	0.01	1.09	0.00	0.00
Al(OH)3	0.00		0.01	0.00	0.00	0.00	- 0.06

Figure 12 Sample of input data to HSC Sim Recycling simulation models as derived based on data analyses and processing (full stoichiometric composition and masses) (converted from MISS to xlsx format and data bases in HSC Sim models)

To high quality product or intermediates for further processing in this model	
Selection of elements in product compounds	% Recovery
Ag (99.999% purity) electrolytic	74.24
Al, Ba, Ca, Fe, Mg, Si (as Al2O3, BaO, CaO, FeOx, SiO2 in slag)	99.00
Au (99.999% purity) electrolytic (see PM-PGM Recovery tab)	99.64
Bi (as oxide in flue dust)	99.19
Cu (99.999% purity) electrolytic	97.41
In (to alloy for further processing)	99.99
Sn (to various intermediates for further processing to recover rest)	28.68
Zn (to flue dust for further processing)	99.59
Pb (to flue dust for further processing)	99.53
Plastics recovered as energy and reductant	
Ni (rest to be recovered from slag)	90.15

Figure 13 Example of output data from recycling simulation models: recycling/recovery rates for all elements and total part/product (xlsx format)

20 tph Module 1 (Mod 1) products	Amount	Unit
Copper Alloy	1.16	tph
Multi-Metal Alloy	1.28	tph
Energy (if 30% efficient) Ox	6892.1	kW
Energy (if 30% efficient) Red	1760.6	kW
Per tonne of feed	0.43	MWh/t
Scope 1 GWP	1.55	kg CO2/kg Mod 1
Scope 1 AP (SOx+NOx)	Low	kg SOx-eq/kg Mod 1
Slag (building material)	9.39	tph

Figure 14 Example of output data from recycling simulation models: mass flows of product and emission/loss flows, energy balance/usage and LCA data such as GWP, SOx+NOx (in xlsx format)

3.2.4 Technical Manual

No technical manual is available, as this component is a persistent data store which doesn't need to be installed, configured or directly used by end users.

3.2.5 Licensing

The Data Lake is licensed as a closed-source component with proprietary license.

3.3 Conclusions and Future plans

This chapter concludes the section dedicated to the TREASURE Data Lake, with the conclusions and some considerations about future plans for this component. To this extent, future planned activities are focused on increasing the amount of information at disposal taken from all data sources, in order to obtain a more complete set of information to be exploited by the other platform tools and actors. Concerning the integration activities, those will take place in with other parts of the platform, in accordance with the individual components' development roadmap. Effort on these activities will be also shared with tasks T4.4 and T4.5.

Concerning the next version of Data Lake, it is expected to provide improved capabilities in term of offered data amount to other parts of the platform. This will be possible thanks to the first execution of the platform components that will bring more data into the Data Lake. This improvement process will be reported in the second iteration of this deliverable (D4.4) due to end of M30.

Finally, with respect to the deployment into production, such activity will be delegated to the final version of the data lake that will be described in D4.4; as for the current version of the component, it will be executed inside a development environment for continuous improvement, integration and testing, until the definitive production-ready version of it will be put in place.

4 Data Importer

This chapter focuses on the description of the Data Importer platform component. The section starts summarising the overall information about the software released (description, overall data, functionalities and architecture), after that, technical information is reported about architectural stack, technical manual for installation and licensing (including third parties' components). Finally, the conclusions and future steps closes the chapter.

4.1 Component Description

4.1.1 Overall Data

Item	Value
Component Name	Data Importer
Software version	V 1.0.0
Reference workpackage	WP4
Responsible Partner	TXT
Contact person	Mattia Calabresi, mattia.calabresi@txtgroup.com
Source control	The repository for this component is marked as private
Short Description	The Data Importer is a utility component responsible for connecting to external data sources, retrieve the desired information, convert them into a suitable format and store them in the TREASURE Data Lake

4.1.2 Purpose of the tool

The Data Importer component makes it possible to integrate data coming from sources that are external to the scope of TREASURE project, into the centralized Data Lake, by providing an ingestion pipeline that aims at normalizing the format of external data with the data model adopted by the Data Lake. The normalization process that is performed highly depends on the source and target data format and, therefore, several different conversion procedures are exposed by this component.

4.1.3 Summary of Functionalities

In this section are reported the main functionalities covered by the Data Importer, along with a brief description of the services offered. As such, the most relevant features are the following:

- The ability to connect to designated external data sources through the usage of official dedicated APIs, to retrieve total or partial information included in the data stores.
- The ability to convert the retrieved information into suitable data types to be included in the Data Lake, depending on the format of data from the source and the target data stores. In particular, data are converted in JSON format for non-structured data to be included in the Mongo database, while SQL objects are used for data that has to be stored in relational tabular format.

4.2 Technical Information

4.2.1 Internal Architecture

The technical architecture for the Data Importer component follows a standardized approach, based on development best practices and state of the art technologies. The component is realized as a standalone backend python service that offers different conversion capabilities by exposing dedicated APIs for each external data source and conversion required. The component is made available to the platform in the form of a containerized solution that is hosted, once again, exploiting the high capabilities of the AWS infrastructure. More specifically, the technologies used for the deployment of the Data Importer leverage Amazon ECR and Amazon ECS for a secure and scalable deployment architecture. Amazon Elastic Container Service (Amazon ECS) allows for an easily deploy of containerized workloads on AWS, while Amazon Elastic Container Registry (ECR) offers secure and reliable storage of containerized solutions to be used in conjunction with container orchestration services such as Amazon ECS.

4.2.2 Technological stack

Item	Value
Nature	Backend ingestion pipeline
Programming Language	Python
Development Tools	Visual Studio Code
Additional Libraries	No additional libraries required
Application Server	This is a standalone backend component which relies on a dedicated python server
Databases	This component operates on the TREASURE Data Lake, which is constituted by a series of data sources (including databases)

4.2.3 Technical Manual

No technical manual available, as this is a backend component not exposed to the end user.

4.2.4 Licensing

The Data Importer is licensed as a closed-source component with proprietary license.

4.3 Conclusions and Future plans

This chapter concludes the discussion on the Data Importer by summarizing the future plans expected for the component and providing final thoughts and conclusions. Regarding the next actions to be performed, those will be focused on improving the integration with external data sources whenever needed, in order to continuously develop a tighter integration with those.

As for the integration activities, they will proceed in accordance with the development of other parts of the platform, especially with the data lake as it represents the main utilizer of the data importer. Effort on this activity will also be shared with tasks T4.4 and T4.5. To this extent, the continuous improvement process, as well as the updates on the integration status, will be reported in the second iteration of this deliverable (D4.4) due to end of M30.

Finally, concerning the deployment into production, this activity will be delegated to the final version of the Data Importer that will be described in D4.4. As for the current version of the component, it will be executed inside a development environment for continuous improvement, integration and testing, as already mentioned for the Data Lake component.

5 Conclusions and Next Steps

The present deliverable documents TREASURE Data Lake as a result of the outputs regarding technical requirements and specifications as pinpointed in D4.1, deriving not only from the preliminary system analysis but also from the discussion with the industrial use cases. A complete description of platform use in the three modules is presented, followed by a comprehensive depiction of each component with additional details related to features, functionalities in the selected modules, met requirements and synergy with other key elements.

The next steps will be mainly focused on the execution of the other tasks foreseen in WP4, assigned to technical design, implementation, and integration of the modules. The activities will be carried out strongly collaborating with the other Work Packages, mainly WP3, aimed at providing the main contents of the modules; WP5, aimed at providing a testing simulation, and WP6, aimed at providing a final validation of the TREASURE platform and its technical implementation. The final version of the Data Lake will be provided in D4.4 due on M30 as a result of the refinement of technical developments concerning all platform components, in addition to emerging end user needs and requirements.