PROJECT DELIVERABLE REPORT

Grant Agreement Number: 101058732

Joint Industrial Data Exchange Platform

Type: R

D3.1- Report on iTelos methodology for data search, sharing and interoperability

Issuing partner	UNITN
Participating partners	UCAM, TVS
Document name and revision	D3.1- Report on iTelos methodology for data search, sharing and interoperability
Author(s)	Simone Bocca
Deliverable due date	2024-February-01
Actual submission date	

Project Coordinator	Vorarlberg University of Applied Sciences
Tel	+43 (0) 5572 792 7128
E-mail	florian.maurer@fhv.at
Project website address	www.jidep.eu

Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission services)	
CO	Confidential, only for members of the consortium (including the Commission services)	
SEN	Sensitive, limited under the conditions of the Grant Agreement	✓

Content

1. Introduction

1.1 Executive Summary

This is the second, and final, version of a report describing the iTelos methodology for building reusable purpose-specific Knowledge Graphs. This document is based on the previous version D3.1- Report on iTelos methodology for data search, sharing and interoperability (Beta). In the context of the JIDEP European project, iTelos is applied to produce high quality reusable data, representing products, materials, composite materials as well as electrical and mechanical components to be used or recycled by the final users as described by the project use cases. The current report is focused on the definition of the methodology, by describing its approach focused on the purpose for which the KG has to be built, as well as the methodology phases with their inputs, outputs and the activities performed within them. The next section of the report indicates the intended audience for this report, while in Section 2 the methodology is better detailed. More precisely, Section 2.1 describes the iTleos methodology approach, while the sections from 2.2 to 2.6 describe the different methodology phases. Section 2.7 describes the characteristics of the iTleos outcome. Section 3 briefly describes the knowledge modelling approach adopted by iTelos to produce high quality shareable schema resources. Then Section 4 briefly describes the Open Data (OD) environment that supports the methodology, with the objective to enhance the reusability of high-quality resources. Section 5 reports some application examples of the iTelos methodology over the three JIDEP use cases. In the end, Section 6 concludes the report with final observations and the aspects that will be considered in the next version of the report.

1.2 Intended audience

The intended audience for this report is composed by the following subjects:

- Project's domain experts: subjects involved in the project, having major competencies in the domain of interest considered by the JIDEP project. These subjects are involved in the data production process, following the iTelos methodology, with the objective of supporting more technical roles during the data and schema modelling.
- Data scientists: technical subjects involved in the iTelos methodology process, by acting over the data management activities (data collection, cleaning and formatting).
- Knowledge engineers: technical subjects involved in the iTelos methodology process, by acting over the knowledge management activities (reference ontology collection, schema modelling, schema alignment).

1.3 Updates with respect to previous version

This deliverable is the updated (final) version of the first (beta) version of *D3.1 - Report on iTelos methodology for data search, sharing and interoperability*. The difference with respect to the previous version of this deliverable, is the addition of a

dedicated activity to the methodology, called Purpose Formalisation. Such an activity improves the methodology, by structuring a concrete process for the definition of the functional and non-functional requirements for the KG to be produced in output. The impact (concretely, the phase's output) that this activity has on the methodology phases is described in each phase related section. Moreover, the final version of the deliverable describes how the methodology has been applied on the three different use cases, defined by the JIDEP project.

2. The iTelos methodology

Interoperability is a key feature for the composite material data handled, produced and shared by the JIDEP project. Interoperable data must be in turn, findable, retrievable, integrable with other data as well as shareable with the objective to be available for other applications. To achieve such an objective the JIDEP project adopts the iTelos methodology whose objective is to produce data strongly based on such crucial data quality aspects. To this end iTelos is designed to handle and produce (or transform existing data in) Knowledge Graph (KG) resources. This choice has been made considering the nowadays large adoption of KGs, thanks to their scalability and adaptability to several applications and application domains.

2.1 iTelos approach

More in detail the iTelos methodology aims at producing interoperable purpose-specific KGs. The iTelos approach is depicted in Figure 1. The methodology logical sequence, represented by the dashed lines, shows the user which provides an informal specification of the problem she wants to solve, what is called *The Purpose*, and receives in output a KG, called, in Figure 1, the *Entity Graph*. The concrete factual process, implementing the methodology, is represented by the four solid lines which indicate the reuse of the prior resources, represented at the data level as Datasets and at the schema level as Ontologies. It is important to notice how the iTelos methodology strongly considers the reuse of already existing data and schema level resources. This approach aims to reduce the effort in the production of KGs, as well as to produce resources that are in turn highly shareable, thus enabling a reuse and share data recycle loop supporting the data interoperability.

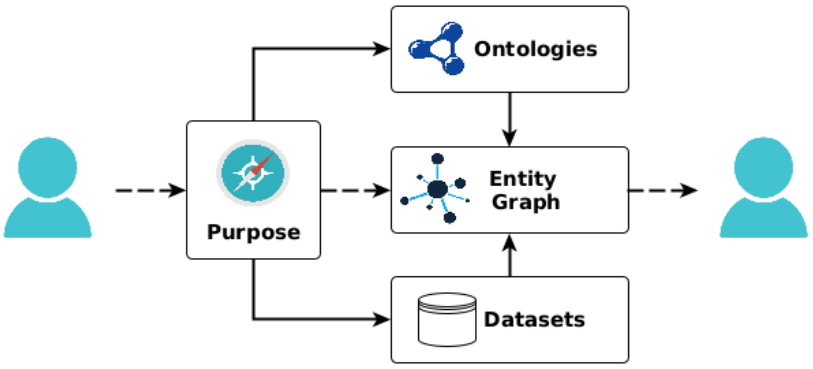

Figure 1. The iTelos Approach

The main input of iTelos is the Purpose provided by the user, which plays a crucial role in building KGs suitable for specific applications, by following the requirements that the user has in mind. ITelos defines the Purpose as a composition of four main elements, as follows:

- A list of functional requirements which the final application must satisfy, by exploiting the KG produced. Such requirements are concretely represented by a list of competence questions. Nevertheless, the final users, namely the people that are interested in the application which exploits the KG produced, are not always able to define such requirements in the form of a set of (well-structured) competency questions. This is because, often, the people who require the information carried by the KG, are not the same people who are in charge of building that. As a consequence, the final users could not have the necessary expertise to structurally define the initial requirements in a way that they will be suitable to lead the KG generation process. For this, the iTelos methodology considers the first component of the input Purpose, as a natural language statement, through which the final user expresses the final goal that the KG should be able to satisfy.

- A set of datasets to be reused (and integrated), collecting the information to be included in the final KG. Such datasets can be directly provided by the user, or even collected from external data sources. The key aspect, here encapsulated in the Purpose, is to consider the reuse of an already existing dataset, starting from the beginning of the process. For this reason, the initial set of datasets, included in the Purpose, strongly affects the effort required for building the final KG. The more datasets are considered, the less effort will be required, during the process, to satisfy the initial requirements, as well as a higher level of precision (in the requirements satisfaction) of the final KG will be reached by the process.

- A set of pre-existing reference schemas, ontologies but not only, whose reuse will facilitate the development of common schema parts, to be included in the final KG, which can be shared by future applications. The key points, here introduced by iTelos, is the possibility to find the right level of interoperability between the KGs produced by the process, by collecting well-known and widely used schemas to represent common information across applications and application domains. Such reference schema can be both directly extracted from (or sometimes, provided with) the dataset mentioned previously, and collected from existing schema repositories; for instance, LOV, LOV4IoT, and DATAHUB, are three among the most relevant repositories, which collectively contain around 1000 ontologies, some of which contain thousands of elements.
- A set of metrics, whose goal is to evaluate the three key elements of the purpose, competence queries, reference schemas and data, in terms of use and reuse. More precisely, the evaluations metrics aim at ensuring that the

resources handled along the iTelos process, maintain a certain level of suitability, to satisfy the user specific requirements, as well as to be reused for different applications and application purposes.

A crucial design decision in the structure of the Purpose, which reflects into the overall iTelos process, is that the data level and the schema level are kept distinct and independent. This assumption is the key, as it allows to split the problem of reusing existing data from the problem of facilitating the shareability and future reuse of the KG being developed. More in detail, along the iTelos methodology process, both the data and schema resources are handled separately as KGs, and only in the end they are merged into a single final KG, that in Figure 1 is called Entity Graph (EG).

At data level the resources are handled as EGs themself, but unlike the final EG, they are treated singularly, dataset by dataset, thus still not integrated with each other. EGs are graphs where nodes are *entities* (e.g., my cat Garfield), decorated with data property values describing them, and where links are object properties which describe the relations holding between any two such entities.

At schema level the resources are KGs that we call Entity Type (*etype*) Graphs (ETGs), namely KGs which define the schema of EGs. In other words, for each EG there is a corresponding ETG which defines its schema. In ETGs nodes are etypes, namely classes of entities (e.g., the class cat), each described by a set of data properties and by a set of object properties which define the range of links allowed among the nodes of the EG defined by the ETG. Datasets and ontologies in Figure 1 are, respectively, examples of EGs and ETGs.

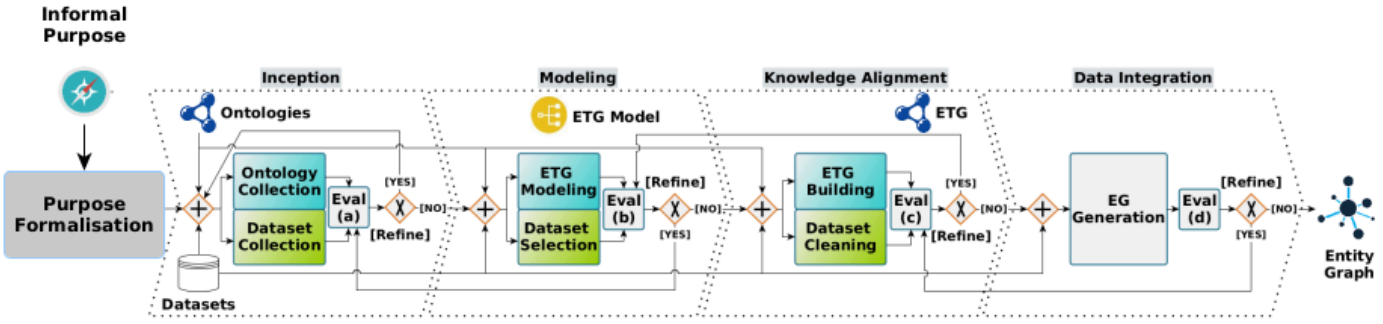

Figure 2. iTelos Process

Considering the Purpose, described above, as main input of the methodology process, and the purpose-specific KG as main outcome, the implementation of the process is defined over four phases plus a crucial initial activity, as graphically described in Figure 2 and initially described as follows:

1. Purpose Formalisation: this activity has the objective of formalising the initial, non-structured, version of the final user Purpose. As already anticipated, the users usually express their objectives as natural language statements. This

initial activity aims at extracting from that statement all the possible information elements (important concepts and terms, as well as, the real world entities involved). The output of such activity is a formalised version of the purpose that will be exploited to build the KG in the following phases.

2. Inception: the first phase takes in input the user's purpose, previously formalised, to the extent that it is needed, into a refined set of Competency Queries (CQs). Moreover, as part of the phase input, there is a first validated collection of input datasets and input ontologies. During this phase all the three elements of the Purpose are used to collect the most accurate set of information to be handled and, finally, included in the KG.
3. Modelling: the second phase builds a model of the ETG to be used to represent the information of the final KG. Such a model is built by taking into account all the three elements of the input Purpose (formalised CQs, datasets and ontologies).
4. Knowledge Alignment: the third phase builds the ETG based on the model designed previously. The key idea, in this phase, is to build the most shareable ETG via the reuse of the selected reference ontologies.
5. Data Integration: finally, the last phase builds the output EG by integrating the input datasets into the ETG.

As depicted in Figure 2, at the end of each phase an evaluation activity is present. Such activity aims at checking if the intermediate outcomes of the phase just completed, have been properly produced, thus satisfying the right levels of use (over the purpose specific requirements) and reuse (over the shareability requirements). In the next sections of this report a more detailed description of the iTelos phases is provided.

The iTelos process is supported by a dedicated Open Data environment providing quality, already existing, data and schema resources, which can be reused to satisfy the requirements of the Purpose in exam. The Section 4 of this report provides a description of such a supporting infrastructure, however, more details are reported in the deliverable D3.3.

2.2 Purpose Formalisation

The first activity of the iTelos methodology aims at formalising the initial informal version of the Purpose, as it has been expressed by the final user. As already anticipated above, due to the lack of expertise the users usually provide in input the definition of their objectives represented as a statement expressed in a natural language. It is difficult to extract the functional, and non-functional, requirements from this initial version of the user's main objective. The current activity is composed of four steps having the objective of identifying all the possible information elements that, in the beginning, are implicitly defined within the informal Purpose statement. Here below the four activity steps are detailed, thus defining the execution of the Purpose Formalisation activity.

- Step 1 - Personas & use cases definition: this activity aims at identifying the background information, by defining formally which are the main elements composing the user main Purpose. To this end the user (assuming that the user is the domain expert relative to the Purpose she aims to achieve) is asked to define three main elements, implicitly defined into the initial, informal, Purpose. Such elements are described as follows, as well as depicted in figure 3:

    - *The Domain of Interest*: It refers to the area of knowledge, or field of study, of interest. Examples are the domains capturing knowledge about daily lives, such as music, tourism, and health, or geographical domains, like a specific region or even a state or a political space like the European Community.

    - *The Context*: It refers to the description of a specific context existing into the domain of interest. The first prescriptive definition of Context referred to it as a location, identities of nearby people and objects, and changes to those objects . More in details the context is defined over three main dimensions:
        - Geographical boundaries: Aspects that geographically constrain the problem.
        - Temporal boundaries: Aspects that constrain the problem in time.
        - Domain boundaries: Domain specific aspect constraining the problem.

    - *The Personas and use cases*: they are user-centred subsets triggered by various subjects, called Personas, and their real-world actions into a specific context, called Use Cases (or Scenarios). Personas generation is a widely heralded technique that provides semi-fictional subjects characterising the perception and needs of larger groups of end-users. Moreover, Use Cases are an essential complement to personas, ensuring a complete and good representation of end-users.

  The three elements described above, clearly define how the user intends to exploit the KG that has to be created. After this initial step the Purpose background information is clearly defined, thus allowing to extract more precise information for the definition of the functional and non-functional requirements of the final KG.

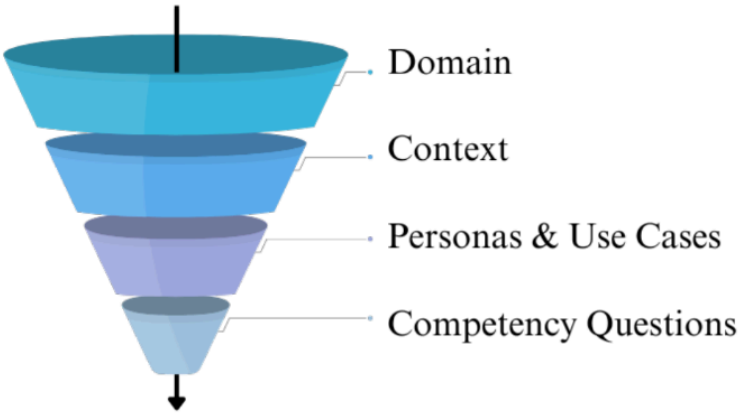

Figure 3 - The Purpose information elements

- Step 2 - Competency Question definition: the second step adds one more step in the formalisation of the initial purpose, by extracting the KG functional requirements, from the output of the previous activity, shaping them as Competency Questions (CQs). Formally the CQs can be defined as follows;
    - *A list of natural language questions. Each question defines a need (or query to the KG) that should be satisfied by the final KG. Each query refers to a Persona into a specific Scenario.*

  It is important to notice how the definition of CQs is crucial for the design of the final KG. A poor set of CQs doesn't provide enough information regarding which (information) entities need to be modelled in the KG. A set of CQs with low heterogeneity, does not represent precisely all the possible (information) details that the KG should be able to support. AS

- Step 3 - Concept Identification: having the CQs listed down, the next step aims at extracting the concepts identifying the information entities (and their properties) to be modelled in the KG. The concepts identified at this step are crucial to increase the level of interoperability of the final KG. In fact, once identified such a concept can be aligned with standard terminology (relative to the domain of interest). In this way the standard terms will be used to represent the information concepts into the final KG, thus supporting the future reuse.

- Step 4 - Concept classification: the final step is the direct consequence of the previous one. The idea is to classify the concepts, with the objective of checking which are the most purpose specific, thus representing information that is focused on the specific KG's requirements; and which are the most popular, thus representing information that can be more reusable (both considering the resources to be collected to build the KG, and the final KG itself which can be more reusable if more reusable are the concepts used by its data). To this end the classification is defined over two criteria:

- the *Focus*, which defines how much an entity is "important" respect to the main purpose;
- the *Popularity*, which defines how much an entity is reused in already existing data (considering the input information sources).

Both Focus and Popularity, for each entity, can have three value:

- Common: (Focus) general entities for the purpose considered. (Popularity) the entity is largely available in existing resources.
- Core: (Focus) specific entities for the purpose considered. (Popularity) the entity is available in existing resources but not so common.
- Contextual: (Focus) very specific entities for the purpose considered. (Popularity) the entity is not available in existing resources.

2.3 Inception

The Inception phase takes as input the Purpose, with all its components (formalised CQs and concept extracted, plus the datasets and ontologies provided initially), and concretely collects the full set of ontologies and datasets needed to build the target EG.

To better describe the iTelos methodology, we consider, as input Purpose, the JIDEP project's objective as the need to create a Knowledge Graph with material passports of products and components, where raw materials, such as a certain quantity of carbon fibre extracted from an automotive monocoque, being considered as products. Considering this objective, the following input data were collected from the automotive data-providing partners, and schema resources were created by the data processing and harmonisation partners in T2.1.

The input *data resources* consist of 800 grams of carbon epoxy, 352 grams of polyurethane, 56 grams of glass epoxy and 350 grams of aluminium in a cross beam of monocoque. The monocoque has eight equal dimension and weight cross beams, four on the right-hand side, 1RH-4RH and four on the left-hand side, and 1LH-4LH, supplied by ADL partner.

The input *schema resources* include the material passport properties defined to describe products, components and their constituent materials to enable the development and publishing of material passports, as well as to develop a material circularity calculator to promote a circular economy. The material passport includes the identification properties of products and components, such as name, brand/trade name, manufacturer details, the global trade item number (GTIN) or European article number (EAN), functionality and image. It also covers physical, temporal, thermal, biological, temporal, and compositional properties. The following properties of products and components are included in their material passports:

1. Identification Properties
   a. Name
   b. Level (It can be 1, 2, 3, etc. For example, while providing input about monocoque, which can be represented as a product, the user will assign 1, but for cross beam 1RH, which is a component of monocoque, the user will assign 2)
   c. Part of (A component can be part of another product/component. For example, cross beam 1RH is part of monocoque)
   d. Trade name
   e. Brand name
   f. Manufacturer
      i. Manufacturer name
      ii. Registration number
      iii. Registration country
   g. GTIN
   h. EAN
   i. Functionality
   j. Automatic tracking/scanning
   k. Image
      i. URL

2. Physical Properties
   a. Density [g/cm-3]
   b. Dimension
      1. Height [cm]
      2. Width [cm]
      3. Length [cm]
   c. Resistance
      1. Compressive strength [Pa]
      2. Shear strength [Pa]
      3. Tensile strength [Pa]
   d. Rigidity
      1. Shear modulus [Pa]
      2. Young’s modulus [Pa]
   e. Mass [g]

3. Thermal Properties
   a. Heat transfer coefficient
   b. Thermal conductivity [W/(m-K)]

4. Temporal Properties

   a. Expected lifetime [y]
   b. Service life [y]

5. Biological Properties
   a. Biodegradability
   b. Decomposability

6. Compositional Properties
   a. Chemical composition
   b. Ingredient [g]
   c. Recycled content [g]

Notice how the above properties are defined by using the concepts that have been extracted in the previous Purpose formalisation activity. Such concepts were implicitly included in the requirement described by the following competency question: "*How to identify all the materials that compose a monocoque, in the automotive sector?*". The CQs generated are then matched to the input datasets and ontologies provided as part of the Purpose, thus implementing the activities of Ontology Collection and Dataset Collection, as from Figure 2. Depending on the process, the final list of collected datasets and ontologies may extend, and partially overlap with, the input list. In other words, the formalised list of CQs generated, leads to the collection of the resources required to build the final EG. The key observation is that matching CQs, the ontologies and (the schemas of the) datasets is crucial for the success of the project. A little coverage would mean that there is not enough data for the implementation of the EG and that there should be a revision of the CQs. For this reason, multiple iterations of the Inception phase are considered before reaching a set of resources (both at schema and data level) suitable to satisfy the requirements defined, and then proceed to the next methodology phase. The first intermediate evaluation activity, at the end of the inception phase, has the objective to check that kind of satisfiability, by measuring how much the datasets and ontologies collected can cover the entities and properties identified in the CQs.

2.4 Modelling

The Modelling phase receives in input the ontologies and datasets previously collected, as well as the CQ list plus the list of Purpose's information concepts. The main objective of this phase is to build the most suitable model of the ETG to be used as the schema of the final EG, which in Figure 2 is called the ETG model. In practice, the ETG model includes all the etypes and properties (represented by the Purpose's information concepts) needed to represent the information required by each CQ, possibly extended by extra etypes and properties suggested by the datasets. This extension is optional but suggested to allow for future expansions,

given that the availability of data would make this step low cost, in particular, if considered since the early stages. Based on the theory described in [2] as well as described in the JIDEP deliverable D2.3 - Refined and extended domain-specific ontologies - the ETG model is a teleology that structurally models the requirements extracted from the Purpose, as etypes and properties.

A portion of the teleology built in this phase, for the same example started in the previous phase, is described here below. It represents the Materials Passport which is a knowledge resource developed by the University of Cambridge (UCAM) in collaboration with Technovative Solutions (TVS) with a specific purpose to describe products, components and their constituent materials. It enables the development and publishing of material passports to develop a material circularity calculator to promote a circular economy.

The classes encoded in the materials passport teleology describe, amongst others, Biodegradability, Biological property, Chemical composition, Component, Compositional property, etc. The object properties it encodes include, for example, Empirical formula, has value, has Unit, Image. Finally, some of the data properties it encodes are Automatic tracking/scanning, Brand name, EAN, Functionality and GTIN.

In the current phase, the main objective of ETG Modelling activity is to compose the etypes and properties, extracted form CQs (Purpose's information concepts), into an EER model representing the teleology. Essentially, this activity adds another step to the sub process (started in the Purpose Formalisation activity) of transforming the CQ list into an ETG, thus fully formalising the Purpose functional requirements. In parallel, the Dataset Selection activity finalises the selection of those datasets (between those collected in the Inception phase) which are effectively required, pruning away useless data resources. The Data Selection activity is parallelly supported by the ETG Modelling activity, since the definition of the ETG model makes clear the understanding of which datasets (or portions of datasets) are required and which are not. The JIDEP knowledge resources modelling, is briefly described in Section 3, with the aim of giving completeness to the current report, nevertheless the data structures adopted for the modelling activities are better detailed in the dedicated deliverable D2.3 on refined and extended domain-specific ontologies. At the end of this phase, the evaluation activity aims at checking how much the ETG model produced covers the entities and properties extracted from the CQs, as well as if the filtering activity over the datasets collected make the dataset more suitable to satisfy the Purpose requirements. If such criteria are not satisfied (i.e., a certain level of coverage is not reached out), the iTelos methodology process returns to the previous phase with the objective to improve the input of a new Modelling iteration and, because of its execution, its output too.

2.5 Knowledge Alignment

The Knowledge Alignment phase takes as input the ETG model previously generated, plus the filtered set of datasets and reference ontologies. The main objective of this phase is to enhance the shareability of the final EG, by building in turn a shareable and interoperable ETG that fits in the best way possible the datasets to be integrated (represented by the ETG model) as well as the reference ontologies selected. The key aspect in this phase is to consider the input datasets and reference ontologies as prior well-known standardised knowledge. It is important to notice that the input ETG model is itself a possible solution to be considered as ETG. Nevertheless, such a model is too purpose-specific to be exploited (reused) in different contexts (for different purposes), or even to be applied on new data coming for the same context and/or purpose of interest. The alignment of the ETG model with the prior knowledge (ontologies and datasets) builds, in the current phase, an ETG that can be better adapted to other well-defined domains (those implicitly included in the reference ontologies selected in the previous phase) and variations over the dataset to be integrated into the final EG. In other words, the ETG produced is more shareable and reusable, characteristics that will be transmitted to the final EG. As described in deliverable D2.3 on refined and extended domain-specific ontologies - as well as briefly described in this report in Section 3, for completeness, the extension of the ETG model (Teleology) by using etypes and properties of the available reference ontologies, produces the final KG ontology, namely the ETG, that is modelled as a purpose specific, interoperable ontology, called teleontology.

We continue the example reported in the previous phases, by describing how, in the current phase, the ETG model received in input can be merged with the input reference ontologies. We describe below briefly the ETG in the materials modelling domain – the European Materials Modelling Ontology (EMMO) [5, 6], suitable fragments of which can be exploited and reused within the *iTelos* knowledge alignment phase to enhance the reusability and shareability of data resources. EMMO is specific for materials modelling, developed as a multidisciplinary outcome of the European Materials Modelling Council (EMMC). It offers a standard representational framework (an ontology) based on latest materials modelling knowledge, including physical sciences, analytical philosophy and information technologies. The multidisciplinary basis of the EMMO is illustrated by Figure 4, which is extracted from JIDEP deliverable D2.3 on refined and extended domain-specific ontologies - and reported here for completeness in the description of the modelling example. Figure 4 also illustrates how the EMMO interconnects the physical world, materials characterisation world and materials modelling world.

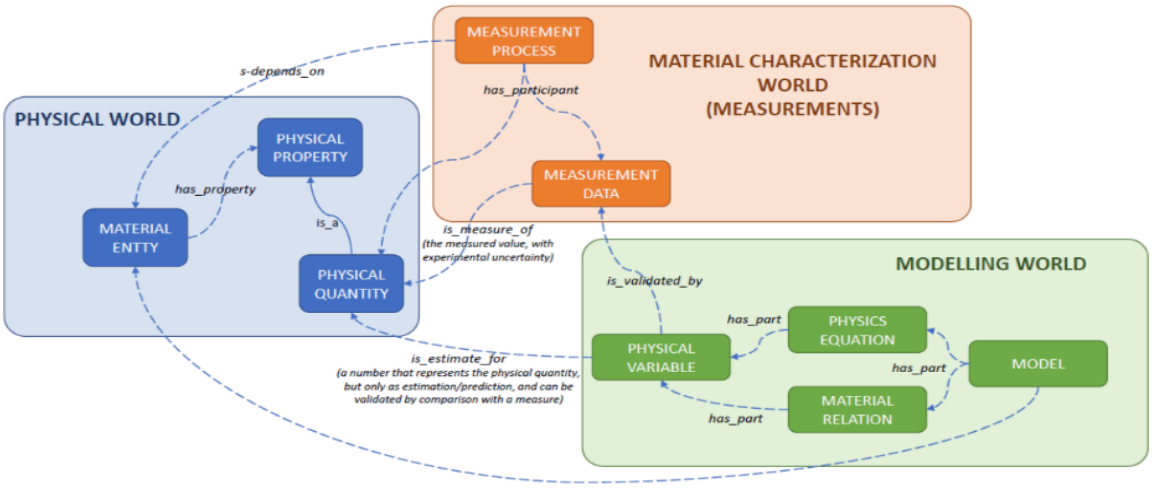

Figure 4. How EMMO interconnects the Physical, Material and Modelling Worlds.

Concretely, the ETG Building activity, visible in Figure 2, is implemented via the Machine Learning algorithm described in [4]. Such an algorithm takes as input the set of reference ontologies and the ETG model and it outputs the final version of the ETG. The resulting ETG is verified for compliance with the input datasets before the final approval. The algorithm aims to identify, from the reference ontologies, which are the best etypes and properties that best fit the datasets and follow the ETG model. Such etypes and properties are used to generate an RDF file, following the OWL language, representing concretely the final ETG. It is important to notice how the final ETG is produced considering the purpose-specific elements (etypes and properties) represented by the ETG model, and the more common and standard elements, represented by the reference ontologies etypes and properties. The two types of elements guarantees that the ETG produced best fits the initial functional requirements as well as the non-functional requirements of shareability and reusability of the final EG, respectively. In parallel the Dataset Cleaning activity performs the final cleaning of the datasets consistently with the ETG. Here, the main objective is the alignment of the data types and formats of the datasets, with the etypes and properties of the ETG. Like in the previous phases the evaluation activity, in the end of the Knowledge Alignment phase, aims to check that the ETG produced can cover the etypes and properties extracted from the CQs, as well as that the shareability level of the ETG complies with the basic constraint. If the evaluation doesn't satisfy the desired criteria, the iTelos methodology process returns to the previous phase with the objective to improve the input of a new Knowledge Alignment iteration and, because of its execution, its output too.

2.6 Data Integration

The last phase is Data Integration. Its input consists of the ETG, and the cleaned datasets previously produced. The objective is to build the EG by integrating the schema and data resources, as adapted to the purpose in the previous phases.

Unlike the previous ones, the data integration phase consists of a single activity, called EG Generation, that merges together the schema and data layers creating the final EG.

To do that, the ETG and datasets are provided in input to a specific data mapping tool, called *KarmaLinker*, which consists of the *Karma* data integration tool [3] extended to perform Natural Language Processing. More details about the Karmalinker tool are provided in the JIDEP tools dedicated deliverable D3.2 (Report on Tools developed for schema and dataset alignment and KG development). Moreover, Karmalinker allows its users to perform minor data format modifications on the datasets, with the objective to align in the best way possible the data to the ETG, hopefully with minimal cost, due to the cleaning operation performed in the previous phase. The first step, in this phase, uses karmalinker to map the data to the etypes and properties of the ETG. The following step is the generation of the entities that are then matched and, whenever they are discovered to be different representations of the same real-world entity, merged. The above process is iteratively executed over the list of datasets selected in the previous phase, processed sequentially. The evaluation activity in phase 4 has the goal of checking that the final EG satisfies the requirements specified by the Purpose, thus by checking that the final EG is able to answer the CQs initially produced. Like in the previous phase a failure in the evaluation activity causes the return to the previous phase. The process concludes with the export of the EG into an RDF file.

2.7 iTelos outcome

The key observation is that the iTelos process is not designed to be bottom-up, thus building the EG starting from the dataset to be integrated, thus resulting too data specific. Neither top-down by considering the modelling of the EG's structure based on the initial requirements only, with the risk to have difficulties in fitting the available datasets with the schema produced. The process is instead a middle-out convergence between data and schema layer, that is maintained within each phase of the methodology described above. During this process, the initial purpose keeps evolving, building the bridge between CQs, datasets and ontologies, to enforce the convergence of this process, and to avoid making costly mistakes. To this end a crucial role is played by the evaluation activities at the end of each phase. It is important to notice that, as from Figure 2, a failure of the evaluation in any of the steps causes the process to go back to the evaluation step of the previous phase. In the extreme case of a major early design mistake, it is possible to go back from phase 4 to phase 1.

The Knowledge Graphs produced by iTelos, are purpose specific being built with the main guidance of the initial user Purpose. Moreover, they are also grounded in standard and well-formed knowledge, thanks to the alignment with the reference ontologies during the third phase. This second feature allows the KGs to be more shareable in the contexts and domains where such ontologies are applied. The shareability leads, in turn, to the reusability of the iTelos outcome. Therefore, if the

reuse of such KGs is suitable for new Purposes, then the iTelos methodology can be applied by using those KGs as input. The quality of resources already produced by the methodology, in terms of data cleaning, formatting as well as over the data schema adopted (high reusability), consistently reduces the cost of producing a KG supporting the new Purpose.

3. Knowledge modelling and reuse

This section aims at providing more details about how the knowledge resources are handled within the iTelos methodology. If the JIDEP deliverable D2.3 - Refined and extended domain-specific ontologies - describes the knowledge structures adopted by the methodology, and how their are composed and interconnected together, in the current section of this report, is described how such knowledge structures are used to produce the ETG (see Section 2.5) of the final KG as main output of the overall iTelos methodology. To this end we describe here a *general, domain-agnostic* step-by-step knowledge modelling methodology by which a purpose-specific teleology and the relative ETG patterns can be generated in practice. Such a knowledge modelling methodology is included in the overall iTelos, parallelly executed with the data layer activities focused on the datasets that will compose the final KG. Before proceeding to elucidate the steps of the process, we emphasise that the process that we propose here are characteristically split between the roles of - *Classificationist* and *Classifier*. The role of the *Classificationist* is to generate the reference background knowledge (reference ontologies) which can be exploited by the *Classifier* to generate the application-specific schema from the domain-specific datasets. The definition of the knowledge modelling methodology follows the step below.

I. *Generate Reference Ontology (a priori)*: The first concrete step in the knowledge modelling methodology aims at instantiating a main reference ontology, called “the UKC” (see, D2.3 - Refined and extended domain-specific ontologies - for the description of such a knowledge resource) which can eventually be customised for any reference domain (e.g., materials modelling domain) - by enriching the UKC with requisite domain-specific concepts. This step is primarily the concern of the *Classificationist*. The universal reference ontology instantiated with the domain specific concepts required, will enable the exploitation of such formalised concepts in the next steps.
II. *Generate Reference Domain Ontology*: Given the generation of a teleology, as described in Section 2.4 above, the second step of the knowledge modelling methodology concentrates on generating the reference domain ontology based on:
    a. Firstly, by building structurally the reference domain ontology for the reference domain (e.g., materials modelling) either from scratch or by reusing and composing concepts from different existing standard ontologies in the same domain.

b. Secondly, by annotating the concepts used to define the reference domain ontology, built as above, by matching them with those added to the universal reference ontology (the UKC).

This step is again primarily the concern of the *Classificationist*.

III. *Ground Teleology Patterns*: The third step of the knowledge modelling methodology implements the building and grounding of the teleology pattern (e.g., teleology for single concepts (etypes), like automobile composite materials) into the (appropriate concepts in the) reused/generated domain ontology (e.g., EMMO). This step is primarily the concern of the *Classificationist.* In this step the modelling methodology is enhancing the reusability of a single portion of information, enabling their independent exploitation and reuse.

IV. *Generate Application Schema (ETG) aligned to Teleology Patterns*: The final step of the knowledge modelling methodology focuses on extracting schemas by aligning them to teleology patterns. This is concretely done by adding/deleting free attributes as required for the specific application specific use case. This step is primarily the concern of the *Classifier.*

The aforementioned steps also assume an input resource, the so called *LiveKnowledge* data catalogue (namely a catalogue for knowledge resources distribution. See Knowledge Catalogue in the next section), considered as support for the overall iTelos methodology, suitably instantiated for a reference domain (e.g materials modelling), which provides a single window interface to search (via metadata), access and potentially reuse, via continuous iteration, of already built ontologies, teleologies and ETGs for modelling one's own ETG for a particular project.

4. iTelos open data environment

The iTelos methodology, as described above, implements the reuse of existing data and schema resources, with the objective to reduce the effort in building KGs, as well as to produce KGs with a high level of shareability, being grounded in well-known standardised data and reference ontologies. Nevertheless, the resources included as part of the initial Purpose, those provided in input by the methodology's users, may not be sufficient to achieve such an objective. For this reason, the iTelos methodology is supported by an Open Data (OD) environment, composed of data catalogues, data repositories and data retrieval services, which aim at providing reusable resources for specific domains and purposes. The OD environment is accessed and directly exploited by the different iTelos phases, when reusable resources are required in the KG building process.

The OD environment is concretely implemented as a system of connected data catalogues, used to publish the data that can be accessed by the iTelos process. The catalogues allow their users to query the data required, as well as to download such data from the repositories in which they are physically stored. Moreover, with the support of specific services, the catalogues allow their users to identify the

resources more suitable for the purpose to be satisfied, using specific input parameters during the data search activity. More in detail the OD environment supports the users along iTelos methodology, by providing different kinds of resources, which are queried and retrieved from different kinds of catalogues. The different catalogues supporting the iTelos process are described as follows:

- **Knowledge catalogue**: this catalogue collects and publishes “knowledge level” resources. This kind of resources are (entire or portions of) data schemas and reference ontologies, representing data of different domains of interest and modelled to satisfy different purposes. Well-defined and standardised reference ontologies are available in this catalogue, to provide support during the KG’s modelling activity, over different common domains (i.e., geography, transportation, composite material, health, and many others).
- **Data catalogue**: the resources collected and published by this catalogue are high quality datasets already produced by other iTelos executions. Such data is the key to reduce the effort in building a KG, due to their capacity to be adapted and composed to support different purpose’s requirements. Notice how the reuse of resources is enabled in iTelos, thanks to this catalogue.
- **Language catalogue**: this catalogue collects and publishes “Language level” resources. This kind of resources support the production of multilingual KGs, when they are exploited by the iTelos process. The multilingual resources produced by iTleos, thanks to the support provided by this catalogue, have a major level of reusability. In other words, they can be exploited in different contexts (as well as different cultures), without the limitations given by the language used to represent the information carried.

As already mentioned above, the iTelos process interacts with the OD catalogues by downloading resources, to be reused to satisfy the purpose to be processed, as well as by publishing on the catalogues the new high-quality resources produced during each execution of the iTelos process. More in details, the OD environment is exploited by the iTleos phases as described below:

- **Inception phase**: during the inception phase, the iTelos process explores the OD catalogues, with the objective to collect reusable, data and schema resources, thus increasing the set of initial datasets and ontologies, provided in input as part of the main Purpose. The Data and Knowledge catalogues are mainly involved in this phase.
- **Modelling phase**: during the modelling phase, the process has the objective to define the model of the ETG. For this reason, the Knowledge catalogue can support the modelling activity by providing more representational options, if the input schema resources are not enough to reach such intermediate outcome.
- **Knowledge Alignment phase**: the third phase of the methodology is supported by the Knowledge catalogue, to identify which are the most suitable reference ontologies to be used for the generation of the ETG, with the

objective to improve its shareability. Moreover, if the information carried by the datasets, as well as the terms used to define etypes and properties, need to be represented in one or many different languages, the Language catalogue provides the resources which can support such an activity.

- **Data Integration phase**: the last phase, of the iTelos process, interacts with the OD environment by publishing new resources in the different catalogues, with the objective of making them available, for future iTelos execution, or any other kind of exploitation. More in detail, the KG produced, as the main outcome, is published (under the permission of the data owner) on the Data catalogue, while its ETG on the Knowledge catalogue respectively. Moreover, if new language resources have been produced during the KG building process, they will be published on the Language catalogue.

More details about the OD environment architecture and features, are provided in the dedicated JIDEP deliverable D3.3 (Report on developed search services).

5. JIDEP Use cases

This section aims at describing the application of the iTelos methodology over the JIDEP use cases defined over three different data and application domains; automotive sector, wind turbine sector and PCB sector.

5.1 Automotive

The first JIDEP use case interests the automotive field. The CRF partner selected a monocoque cross beam dataset to simulate and analyse the manufacturing process of such a component. The dataset includes composite materials information, such as carbon fibre and glass fibre data, used to predict different physical properties. The partner provided data about the constituent materials of eight cross beams of a monocoque. In the dataset, each cross bream is composed of 800 grams of carbon epoxy, 56 grams of glass epoxy, 352 grams of polyurethane, and 350 grams of aluminium. All cross beams are identical and consist of the same materials with a minute variation in the quantity of materials. By applying the iTelos methodology, an ETG has been produced to represent the information of the automotive use case, into a knowledge graph. A portion of the ETG is reported in Figure 5.

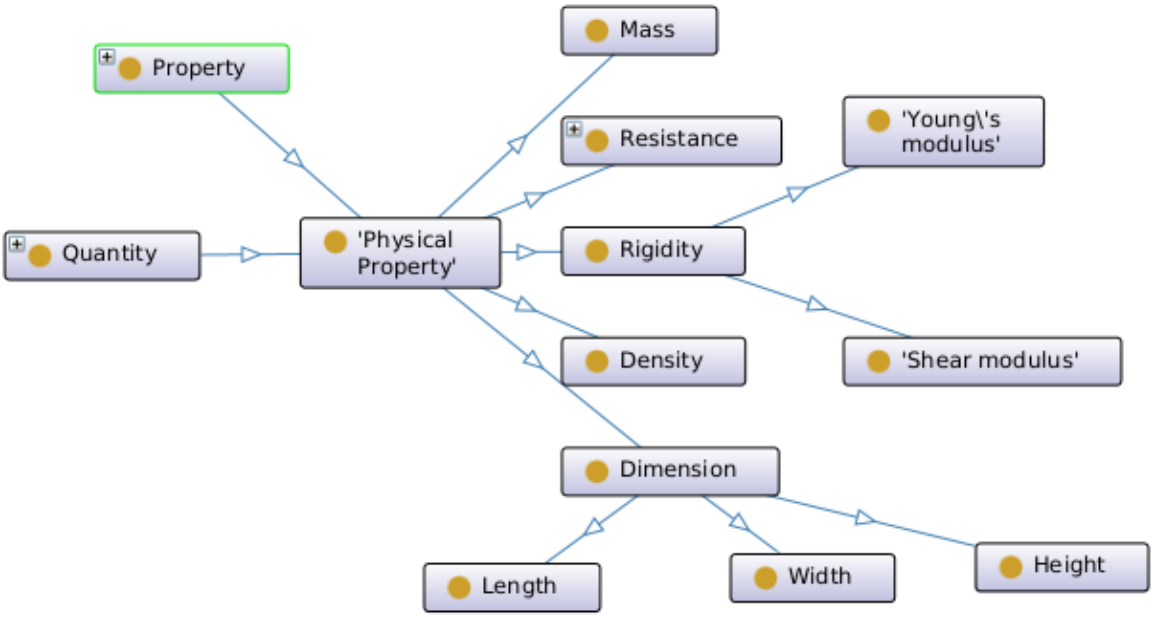

Figure 5 - Portion of automotive use case ETG.

The datasets provided by CRF, have been cleaned and formatted by adopting standard file and data values format. Then, the datasets have been merged with the ETG produced, thus building the final knowledge graph.

5.2 Wind turbine

The second JIDEP use-case interests the wind turbine materials field. The TPI partners provided data about the recycled wind turbine blade used to reclaim glass fibre. The datasets provided include information about the properties of a wind blade. For example its diameter is 163.5 metres, with a root diameter of 3.28 metres, a maximum chord width of 4.2 metres, a maximum laminate thickness of 110 millimetres, a maximum thickness of 110 millimetres, and a total mass of 24,484.4 kilograms. Table 1 shows some of the material data provided for the wind turbine blade.

Blade Material	Amount (%)	Amount (kg)
Balsa core	2.8	688.9
Pet Core	1.5	369.1
Glass Fiber	48.1	11775.0
Carbon pultrusion	15.4	3759.3
Copper	0.7	176.8
Epoxy adhesive	2.6	626.1
Epoxy resin	22.3	5451.0
Steel	5.7	1392.7
Paint	1.0	245.5
Total	100.0	24484.4

Table 1 - Portion of wind blade data.

Figure 6, instead, shows a portion of the ETG for the wind turbine use case, during its modelling by using Protègè, one of the tools included in the iTelos framework described in JIDEP deliverable D3.2, supporting the knowledge modelling activities.

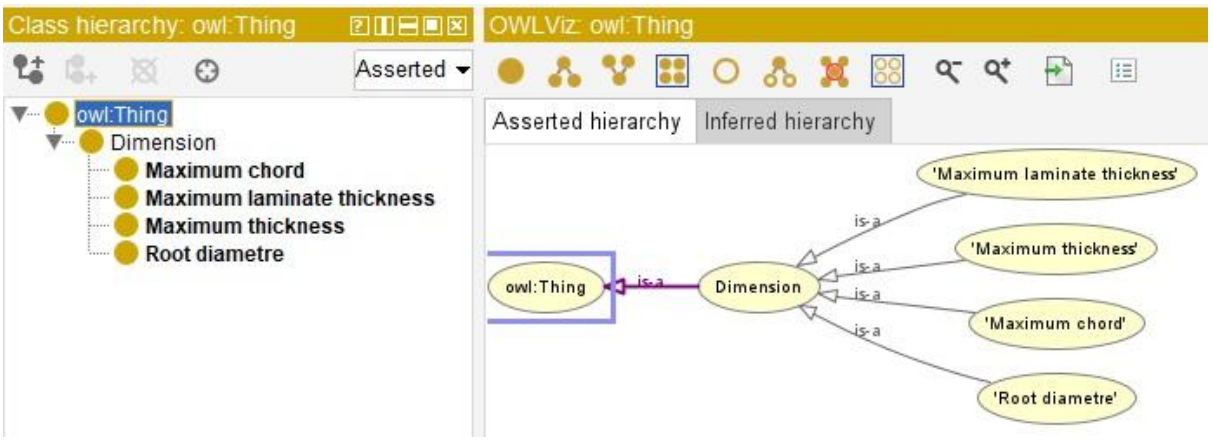

Figure 6 - Portion of wind turbine ETG during the iTelos modelling activities.

Also in this use case, the datasets provided by TPI, have been cleaned and formatted by adopting standard file and data values format, and then merged with the ETG produced, thus building the final knowledge graph.

5.3 PCB

The last JIDEP use case interests the Printed Circuit Board (PCB) The PVI partner is carrying out activities to enable the reuse of whole functioning end-of-life PCBs and electronic components and extract materials from non-functioning PCBs. This partner provided datasets about resistors and capacitors, which are shown in Table 2 and Table 3.

Resistance (Ω)	Power Rating (mW)	Tolerance (%)	Voltage Rating (V)	Operating Temperature Min(C) to Max(C)	Diameter (mm), Length (mm)	Packaging, Termination Style
200	250	5	250	-70 to +130	2.5, 7	Bulk, Axial
2000	500	5	350	-70 to +130	3.8, 10.3	Bulk, Axial

Table 2. A partial dataset describing resistors.

Capacitance (pF)	Voltage Rating DC (V)	Dielectric	Tolerance (%)	Termination Style
0.01	4, 6.3, 10	X5R, X6S	20	SMD/SMT
0.033	16, 25, 50	X7R	1, 10	SMD/SMT

Table 3. A partial dataset describing capacitors.

Figure 7, instead, shows a portion of the ETG for the PCB use case, during its modelling by using Protègè. The datasets provided, after the cleaning and formatting

activities, have been merged with the ETG modelled to produce the final knowledge graph.

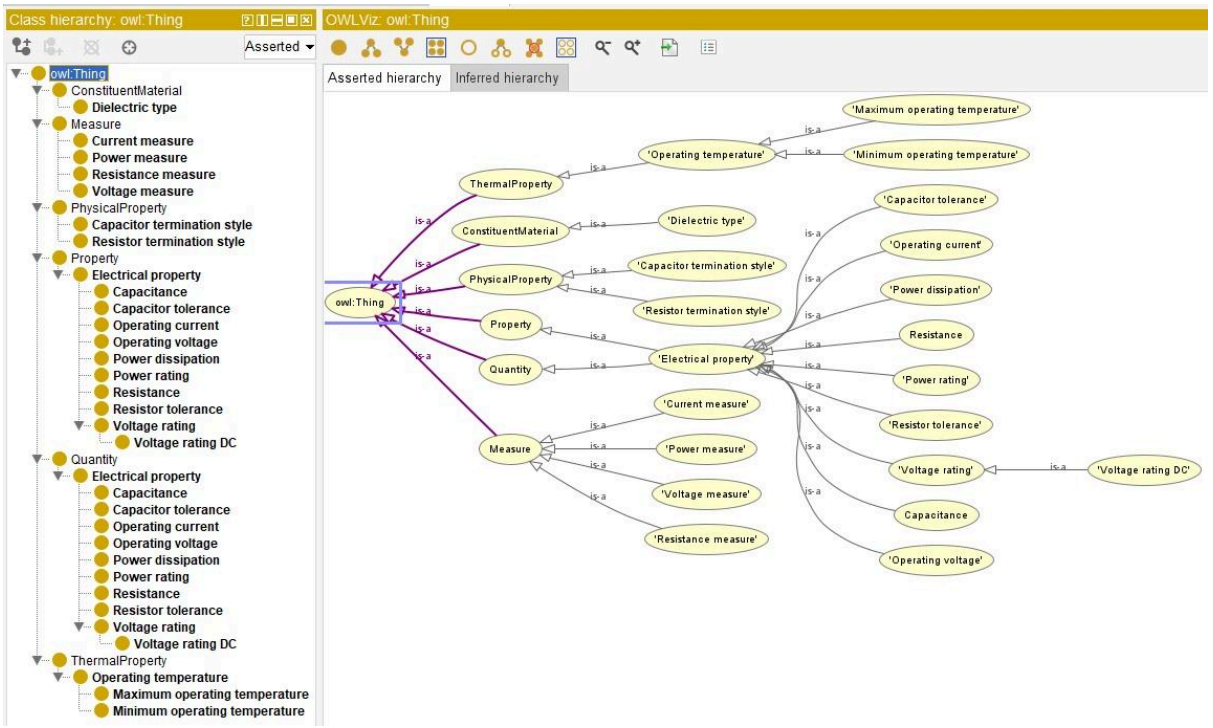

Figure 7 - Portion of the ETG for the PCB use case, during the iTelos modelling activities.

6. Conclusions

The iTelos methodology is adopted in the JIDEP project to produce high quality reusable data, representing products, materials, composite materials as well as electrical and mechanical components (within each specific domain of interest as specified by the project use cases reported in D1.2, Initial Requirements Specification Document) to be used or recycled by the final users identified for the project. Such kinds of data requirements, focused on material and component recycling, implicitly require “recyclable” data that can be easily adapted to different purposes. iTelos fully supports the JIDEP requirements, by implementing the data reuse and share loop.

References

- [1] Giunchiglia, F., Bocca, S., Fumagalli, M., Bagchi, M., & Zamboni, A. (2022, November). Popularity Driven Data Integration. In *Knowledge Graphs and Semantic Web: 4th Iberoamerican Conference and third Indo-American Conference, KGSWC 2022, Madrid, Spain, November 21–23, 2022, Proceedings* (pp. 277-284). Cham: Springer International Publishing.
- [2] Giunchiglia, F., & Fumagalli, M. (2017). Teleologies: Objects, actions and functions. In *Conceptual Modeling: 36th International Conference, ER 2017, Valencia, Spain, November 6–9, 2017, Proceedings 36* (pp. 520-534). Springer International Publishing.
- [3] Gupta, S., Szekely, P., Knoblock, C. A., Goel, A., Taheriyan, M., & Muslea, M. (2015). Karma: A system for mapping structured sources into the semantic web. In *The Semantic Web: ESWC 2012 Satellite Events: ESWC 2012 Satellite Events, Heraklion, Crete, Greece, May 27-31, 2012. Revised Selected Papers* (pp. 430-434). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [4] Giunchiglia, F., & Fumagalli, M. (2020, July). Entity type recognition–dealing with the diversity of knowledge. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning* (Vol. 17, No. 1, pp. 414-423).
- [5] Goldbeck, G., Ghedini, E., Hashibon, A., Schmitz, G. J., & Friis, J. (2019). A reference Language and ontology for materials mode.
- [6] https://github.com/emmo-repo/EMMO

Acronyms and Abbreviations

ADL	ALMAS Partecipazioni Industriali S.P.A.
ADS	Adscensus, MB
AVO	Arteevo Technologies Ltd
BUL	Brunel University London
CRF	Centro Ricerche Fiat Scpa
FHV	Fachhochschule Vorarlberg GMBH
PVI	Precision Varionic International Limited
TPI	TPI Composites
TVS	Technovative Solutions Ltd
UCAM	The Chancellor Masters And Scholars Of the University Of Cambridge
UNITN	University Degli Studi Di Trento
UPCE	Univerzita of Pardubice
ZOREN	Zorlu Enerji Elektrik Uretim As

CFRP	Carbon fiber reinforced plastic
CO2	Carbon dioxide
DLT	Distributed ledger technology
EC	The European Commission
ELV	End-of-life-vehicle
EOL	End-of-life
GW	Giga-Watt
IC	Integrated circuit
Mt	Mega-tons
NMF	Non metallic fraction
PCB	Printed circuit board
R&D	Research & Development
RSD	Requirements specification document
SME	Small-medium enterprise
WEEE	Waste electrical and electronic equipment