PROJECT DELIVERABLE REPORT

Grant Agreement Number: 101058732

Joint Industrial Data Exchange Platform

Type: Deliverable Report

D2.5 Report on ontology-based data documentation

Issuing partner	University of Cambridge (UCAM)
Participating partners	Universita di Trento (UNITN) Technovative Solutions (TVS)
Document name and revision	D2.5 Report on ontology-based data documentation
Author	Dr Md Hanif Seddiqui, Dr Feroz Farazi.
Deliverable due date	29 February 2024
Actual submission date	29 February 2024

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	

Contents

Disclaimer

JIDEP has received funding from the European Commission under the Grant Agreement no.101058732. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Acronyms and Abbreviations

TVS	Technovative Solutions
UCAM	University of Cambridge
UNITN	University of Trento
JIDEP	Joint Industrial Data Exchange Platform
ISO	International Organization for Standardisation
DC	Dublin Core
DCMES	Dublin Core Metadata Element Set
DDI	Data-Driven Innovation
SDMX	Statistical Data and Metadata eXchange
OM	Ontology of Units of Measurement
QC	Quality Check
PCB	Printed Circuit Board
mW	Milli-Watt
C	Celsius
V	Volt
SMD/SMT	Surface Mount Device and Surface Mount Technology
UKC	Universal Knowledge Core
LC	Language Core
CC	Concept Core
EMMO	Elementary Multiperspective Material Ontology
CQs	Competency Questions
ER	Entity Relationship
OWL	Web Ontology Language
RDF	Resource Description Framework
XML	eXtensible Markup Language
SPARQL	SPARQL Protocol and RDF Query Language
SHACL	Shape Constraints Language

1. Executive Summary

This deliverable has a number of explicit goals: (i) to analyse datasets of materials, mainly composite materials and composite material-manufactured products collected in T2.1 and provided by use case partners to extract metadata, (ii) to map the extracted metadata with the existing metadata standards such as Dublin Core and ISO/IEC 11179 (Metadata Registry Standard) to understand the gap between the metadata representation requirements for JIDEP and the available metadata standards, (iii) to develop an ontology for representing metadata required for describing datasets by reusing concepts and properties from the available metadata standards and by creating new concepts and properties to bridge the gap, and (iv) to integrate this ontology with the domain and application ontologies defined in T2.3 and T2.4 to represent data and metadata. To achieve these goals, we have created an ontology for data documentation and tested its capabilities by representing data documentation of all collected datasets, including the ones provided by use-case partners.

2. Introduction

This deliverable is on ontology-based data documentation, which is significant in today's data-driven world. It plays a crucial role in successfully disseminating data collected from different sources (T2.1), including the data-providing partners of the JIDEP project for sharing knowledge, enabling collaboration, and ultimately driving innovation.

In T2.5, which this deliverable is based on, we performed a detailed analysis of datasets focusing on metadata management of materials datasets related to composite materials and their manufactured products. These datasets, obtained through collaborative efforts and contributions from our use case partners, were the foundation for exploring metadata extraction, representation, documentation, and integration within the JIDEP framework. In addition to these datasets, we reviewed the existing standards, ontologies, and scientific papers on ontology-based data documentation. These enabled us to define an ontology for creating ontology-based data documentation by applying a four-pronged approach, as shown in Figure 1.

Metadata Extraction: Datasets collected from use case partners in T2.1 were meticulously examined to delve into their intricacies. We extracted metadata and their associated relationships to capture information about the data origin, content, structure, and usage guidelines. While the data origin accommodates the metadata, such as about, creation, publication, versioning, etc., the content describes different variables, content nature, their datatypes, units of measure, data ranges, and so on. The data structure reflects data format, category, and relationships among data elements/fields. Usage guidelines pertain to organisational and legal interoperable properties, such as policies, terms and conditions, and copyright issues.

Standards Mapping: We aligned the extracted metadata with established standards, such as Dublin Core (DC)[1], Data-Driven Innovation (DDI)[2], Statistical Data and Metadata eXchange (SDMX)[3], Ontology of Units of Measurement (OMU) [1], and Metadata Registry Standard (ISO/IEC 11179)[4]. This process elucidates any gap between the metadata representation requirements for JIDEP and the prevailing standard specifications.

Ontology Development: Recognising the gap, we constructed the data documentation ontology tailored to the specific needs of JIDEP metadata representation requirements. We reused as much as possible from existing standards and created new concepts and properties to fill the gap.

Ontology Integration: We integrated this newly defined data documentation ontology with the domain and application ontologies delineated in D2.3 and D2.4 of JIDEP, forming a holistic and cohesive data and metadata representation system.

The ontology for data documentation is fragmented into several main components: (i) **Data Origin** accommodates several metadata properties, such as title, description, purpose, language, data format, data category, creator, contributor, create-time, update-time, publisher, publishing date, version-no, and version-description; (ii) **Data Content and Structure** describes different variables, their nature such as dependant, independent, and controlled, datatypes and data ranges of data values (content), the characteristic of the content, i.e.,

[1] https://www.dublincore.org/

[2] https://ddi.ac.uk/

[3] https://sdmx.org/

[4] https://www.iso.org/standard/78914.html

qualitative or quantitative, and quantitative data may have unit of measures; (iii) **Quality Check (QC)** related metadata defines data profiling, validation, statistical analysis, and qc tools; (iv) **Usage Guidelines** includes the organisational and legal interoperable properties, such as policies, terms and conditions, copyrights, citation and contacts.

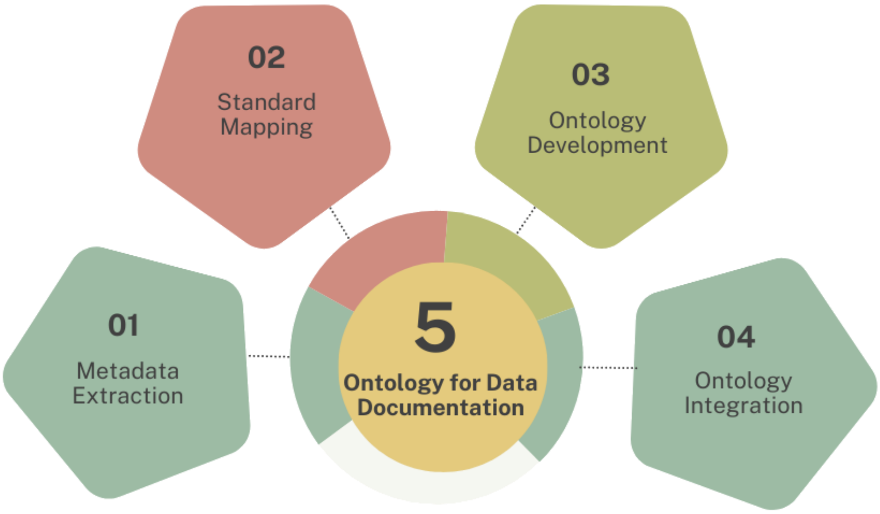


Figure 1. Approach for defining the ontology for data documentation.

The deliverable is organised as follows: Section 3 describes how we analysed data collected from different sources for metadata extraction. Section 4 shows the existing standards we reviewed to map and identify gaps. The ontology development is illustrated in Section 5. Section 6 explains the data documentation based. The quality assessment, i.e., verification and validation processes that ensure the quality of our application ontology, is described in Section 7. Section 8 concludes the deliverable with some future directions.

3. Data Analysis and Metadata Extraction

Datasets provided by the use case partners - automotive, wind turbine, and Printed Circuit Board (PCB) - were analysed to understand their contents, relationships, and internal structure. We studied external parameters, such as data provenance, usage guidelines, and quality check (QC) parameters.

3.1 Use Case Data Analysis

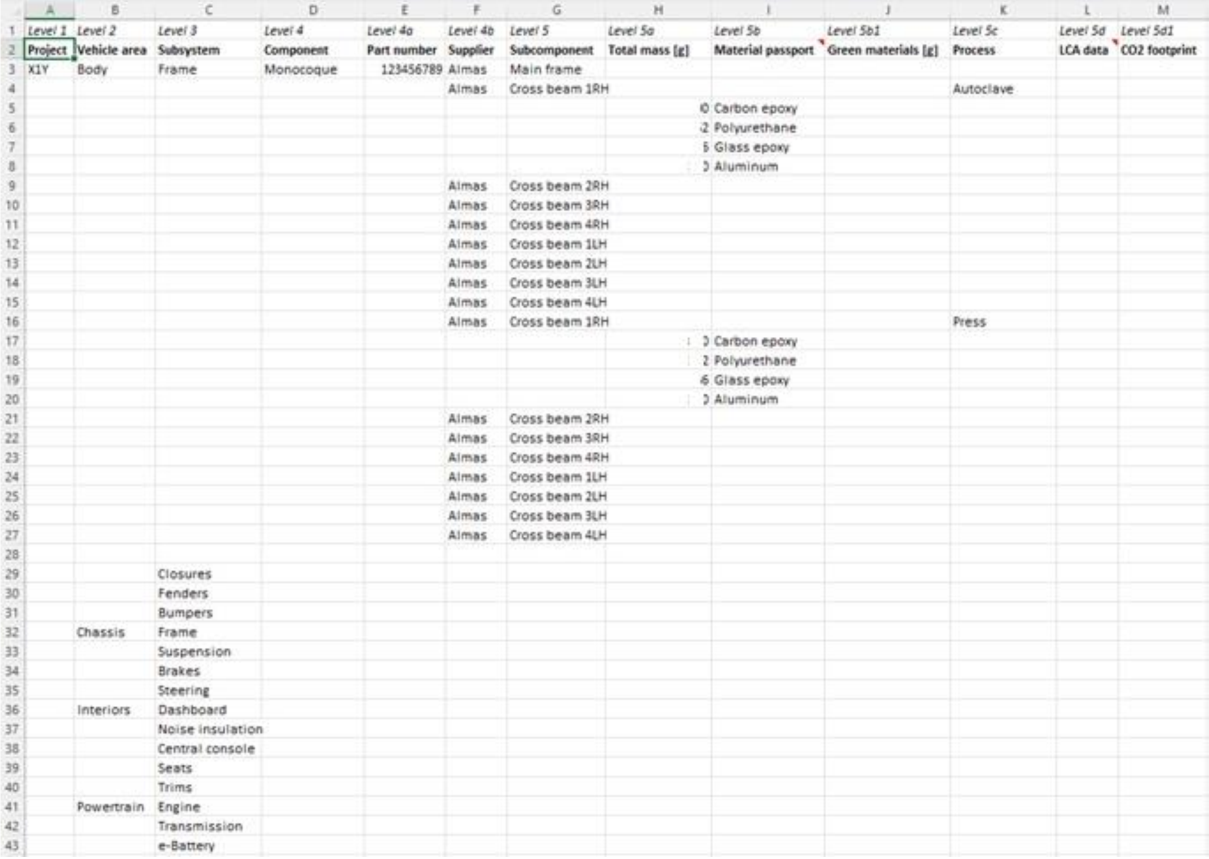
We analysed the internal and external characteristics of data to extract potential metadata. Below, we illustrate the process of data analysis across user cases.

3.1.1 Dataset from Automotive Sector

We received data about automotive cross beams and their manufacturing processes from the automotive use case partner. The dataset describes composite materials in cross beams, such as carbon fibre and glass fibre. The information regarding the constituent materials of the crossbeams was provided. These crossbeams have a consistent composition, albeit with minor discrepancies in material quantities. The dataset was structured hierarchically, encompassing various automotive vehicle parts. The dataset is shown in Table 1.

The eight cross beams were individually crafted within a sealed mould autoclave, adopting a pre-cured shape. Afterwards, the supplier assembles them onto the mainframe chassis using a hand lay-up process. At the same time, the complete chassis component undergoes a concluding autoclave procedure. The composition of the eight cross beams involves four materials: Carbon epoxy, Polyurethane, Glass epoxy, and Aluminium.

Table 1. Details on Manufacturing Process, Materials, and Respective Masses in Automotive Cross-Beam, with intentionally obscured material masses for confidentiality reasons.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	*Level 1*	*Level 2*	*Level 3*	*Level 4*	*Level 4a*	*Level 4b*	*Level 5*	*Level 5a*	*Level 5b*	*Level 5b1*	*Level 5c*	*Level 5d*	*Level 5d1*
2	**Project**	**Vehicle area**	**Subsystem**	**Component**	**Part number**	**Supplier**	**Subcomponent**	**Total mass [g]**	**Material passport**	**Green materials [g]**	**Process**	**LCA data**	**CO2 footprint**
3	X1Y	Body	Frame	Monocoque	123456789	Almas	Main frame						
4						Almas	Cross beam 1RH				Autoclave		
5								0	Carbon epoxy				
6								2	Polyurethane				
7								5	Glass epoxy				
8								0	Aluminum				
9						Almas	Cross beam 2RH						
10						Almas	Cross beam 3RH						
11						Almas	Cross beam 4RH						
12						Almas	Cross beam 1LH						
13						Almas	Cross beam 2LH						
14						Almas	Cross beam 3LH						
15						Almas	Cross beam 4LH						
16						Almas	Cross beam 1RH				Press		
17								0	Carbon epoxy				
18								2	Polyurethane				
19								6	Glass epoxy				
20								0	Aluminum				
21						Almas	Cross beam 2RH						
22						Almas	Cross beam 3RH						
23						Almas	Cross beam 4RH						
24						Almas	Cross beam 1LH						
25						Almas	Cross beam 2LH						
26						Almas	Cross beam 3LH						
27						Almas	Cross beam 4LH						
28													
29			Closures										
30			Fenders										
31			Bumpers										
32		Chassis	Frame										
33			Suspension										
34			Brakes										
35			Steering										
36		Interiors	Dashboard										
37			Noise insulation										
38			Central console										
39			Seats										
40			Trims										
41		Powertrain	Engine										
42			Transmission										
43			e-Battery										

3.1.2 Dataset from Wind-turbine Sector

The use-case partner for the wind turbine shared data regarding the recycled wind turbine blade, specifically designed for reclamation glass fibre. The blade exhibits dimensions: a diameter of 163.5 meters, a root diameter measuring 3.28 meters, a maximum chord width of 4.2 meters, a substantial maximum laminate thickness of 110 millimetres, and an overall thickness of 110 millimetres. The wind turbine blade carries a total mass of 24,484.4 kilograms. The comprehensive material data for the wind turbine blade is presented in Table 2, providing a detailed breakdown of its constituent elements.

Table 2 shows the detailed composition of the wind turbine blade. The intricate breakdown reveals the percentage and corresponding mass of each material used in manufacturing. Notably, materials such as Glass Fiber (48.1%), Carbon pultrusion (15.4%), and Epoxy resin (22.3%) play pivotal roles in shaping the blade's structural integrity.

Table 2. Wind turbine blade dataset.

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

3.1.3 Dataset from PCB Sector

The use-case partner involved in the Printed Circuit Board (PCB) initiative facilitates the reuse of fully functional end-of-life PCBs and electronic components, while extracting valuable materials from non-functioning PCBs. This partner provided data about reclaimed resistors and capacitors, which are described in Table 3 and Table 4, respectively.

Table 3 provides a dataset outlining the characteristics of resistors, including properties such as resistance (Ω), power rating (mW), tolerance (%), voltage rating (V), and the operational temperature range (Min(C) to Max(C)). Additionally, the dataset presents physical dimensions, including diameter and length (mm), along with packaging and termination style details.

Table 3. A partial dataset describing resistors.

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 4 delineates a dataset providing the characteristics of capacitors. Key attributes featured in this dataset include capacitance (pF), voltage rating DC (V), dielectric material (X5R, X6S, X7R), tolerance (%), and termination style.

Table 4. A partial dataset describing capacitors.

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

3.2 JIDEP Ontology Analysis

In the current context, JIDEP possesses a diverse array of ontology facets in the composite materials (D2.3), material passports (D2.4), and application-specific ontologies (D2.4). These ontological frameworks are meticulously analysed to extract metadata. The facets in the composite materials ontology are designed to encapsulate detailed information about the material matrix used, reinforcement type used, manufacturing processes applied, functional requirements fulfilled, etc., fostering a nuanced understanding of their characteristics within industrial applications. The facets of the material passports ontology are tailored to document and track the lifecycle information of products, components and materials. The ontology incorporates metadata related to the origin, manufacturing processes, and subsequent use phases of materials. Furthermore, the application-specific ontologies offer a granular understanding of various applications, such as automotive, wind-turbine and electronic industries.

We analysed concepts, relations, and properties (both datatype and object properties) of these facets to delve into their intricacies. This analysis aims to unveil valuable insights, correlations, and patterns within the data and instances embedded in these facets of JIDEP ontologies.

3.3 Metadata Extraction

Considering all data from use case partners, all facets of ontologies, and external parameters, such as the date of creation and the creator of the datasets, we extracted metadata to delve into not only the intrinsic parameters, including its contents and their intricate relationships but also to address extrinsic factors.

The intrinsic contents and their relations can be represented by different data contents along with their data types. Quantitative and qualitative data can characterise the data, while quantitative data may have units and ranges. Furthermore, data elements can be dependent on other data, independent, and controlled.

Metadata extraction includes different aspects to track the journey of data and understand how it is obtained and transformed, termed as *data provenance* that can also describe the origin, such as source, title, purpose, description, formats, genre, language, creation, contribution and publishing. Metadata may describe the subject and coverage of a dataset, such as product and industry. This information is essential for evaluating fitness and trustworthiness for specific purposes.

Clear and explicit instruction on responsibly and ethically using a dataset while respecting legal and organisational considerations can be described as metadata. Metadata was extracted to include usage guidelines that specify the permission and restrictions on accessibility, usage policies, redistribution policies, citations, and any specific recommendations. Copyright metadata specifies the ownership and usage rights of the dataset for specified purposes and retention policies such as Creative Commons License, all rights reserved, etc. The citation and contact information are also crucial for pinpointing documentation on usage guidelines.

The quality check metadata plays a vital role in establishing the credibility and reliability of the dataset. In this connection, data validation, consistency checking, usability testing, user review, versioning, missing data statistics, etc., were extracted as they enhance the transparency and reliability of the dataset, allowing users to assess the data integrity and suitability for their specific needs.

4. Mapping with Metadata Standards and Ontologies

We carried out a systematic review of metadata standards and a thorough examination of ontological concepts and properties. We mapped the extracted metadata to the vocabulary of metadata standards and ontologies and identified the gaps.

Dublin Core Metadata Element Set (DCMES)

The Dublin Core Metadata Element Set (DCMES)[5] is a fundamental baseline standard in delineating resources within a digital setting. Utilised in the context of ontology-driven data documentation, it provides a framework for structuring and categorising metadata-related attributes. This framework enhances the efficiency of data management and retrieval processes.

[5] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Within the DCMES framework, fundamental metadata elements are defined to capture information about resources. The framework includes title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. These specific elements function as the foundational components, forming the basis for constructing thorough and well-rounded metadata descriptions of resources.

Data Documentation Initiative (DDI)

The Data Documentation Initiative (DDI) is commonly used for data documentation. DDI is an international standard for describing, documenting, and managing data from the social, behavioural, and economic sciences. It provides a comprehensive framework for capturing metadata that describes various aspects of the data lifecycle, including data collection, processing, analysis, and dissemination.

DDI provides rich metadata to describe study design, variables, data structure, sampling methods, data collection instruments, versioning, licensing, confidentiality, restrictions on data usage, data citation, and collaboration. DDI contributes to improved data discoverability, understanding, and reuse, fostering transparency, reproducibility and quality assurance in the research data lifecycle, especially in social sciences.

Statistical Data and Metadata eXchange (SDMX)

SDMX is well-suited for ontology-driven approaches to data documentation, as it provides a robust foundation for representing statistical concepts, classifications, and data structures. Through the application of ontological principles, SDMX facilitates the development of standardised and machine-readable representations for statistical data and metadata. It enhances interoperability and enables automated processes for data integration and analysis.

A notable strength of SDMX in ontology-based data documentation is its capacity to harmonise and exchange data across diverse statistical domains and organisations. By establishing a shared framework for describing statistical concepts and data structures, SDMX facilitates seamless data sharing and comparison. This capability supports cross-domain analysis and decision-making, promoting a more holistic understanding of data trends.

In addition to its interoperability features, SDMX is pivotal in promoting the best metadata management and dissemination practices. It involves capturing and communicating vital information about data sources, methodologies, and quality. The emphasis on transparency, reproducibility, and data quality assurance aligns with essential requirements for informed decision-making and robust policy formulation. SDMX thus serves as a valuable tool in advancing these critical aspects of the data lifecycle.

Ontology of Units of Measurement (OM)

The Ontology of Units of Measurement (OM) plays a crucial role in ontology-driven data documentation, offering a standardised framework for the representation of units of measurement and their interrelationships. OM contributes to precise and semantically enriched descriptions of units, which is crucial for accurately documenting and interpreting quantitative data across various domains.

A significant advantage of OM lies in its capacity to foster interoperability and consistency in data documentation. OM facilitates seamless data integration and exchange across diverse systems and domains through a shared vocabulary and structured representation of measurement units. This ensures a consistent definition and interpretation of units, regardless of the specific context of their application.

Furthermore, OM actively promotes best practices in data documentation by advocating for the adoption of standardised units and quantities. The ontology offers a comprehensive library of predefined units, contributing to the reduction of ambiguity and errors in data representation. This, in turn, leads to enhanced data quality and reliability by providing a solid foundation for ensuring consistency and accuracy in the representation of quantitative information.

ISO/IEC 11179 (Metadata Registry) Standard

The ISO/IEC 11179 standard, also known as the Metadata Registry Standard, provides a framework for defining, managing, and registering metadata within an information system. ISO/IEC 11179 is a foundational tool for organising, describing, and standardising metadata attributes and their relationships when applied within ontology-based data documentation.

This model encompasses components such as data element concepts, descriptions, and value domains. The standard outlines the requirements and guidelines for establishing and maintaining metadata registries. A metadata registry is a centralised repository where metadata definitions, attributes, and relationships are stored and managed. It is a core component for accessing and sharing metadata within an organisation or community.

One of the key strengths of ISO/IEC 11179 for ontology-based data documentation is its emphasis on standardisation and interoperability. The standard defines a common set of metadata concepts, attributes, and relationships, enabling consistent and coherent representation of data elements across different systems and domains. This promotes data integration and exchange, facilitating interoperability between heterogeneous data sources and applications.

While mapping, we found a significant overlap between the extracted metadata describing extrinsic properties and the standards and decided to reuse them. However, we discovered a considerable gap between the extracted metadata describing intrinsic properties and the metadata available in standards. Hence, we decided to create the classes and properties required to represent the intrinsic properties.

5. Ontology Development

5.1 Ontology Development for Data Documentation

We developed OntoDataDoc, an ontology for data documentation to describe data and datasets. The ontology has four facets: data content and structure, data provenance, data usage guidelines, and quality check metadata. The data content and structure facet focuses on data, its characteristics, and its intricate relationships. The data usage guideline facet governs the permissible uses and constraints associated with the data. The data provenance facet focuses on tracing and documenting the origin and history of the data. The quality check metadata facet ensures the integrity and reliability of the extracted data to increase trustworthiness. A dataset may be related to other datasets, and this relationship can be defined via the *hasRelation* property. The bird's-eye view of the whole ontology is depicted **Figure 2**.

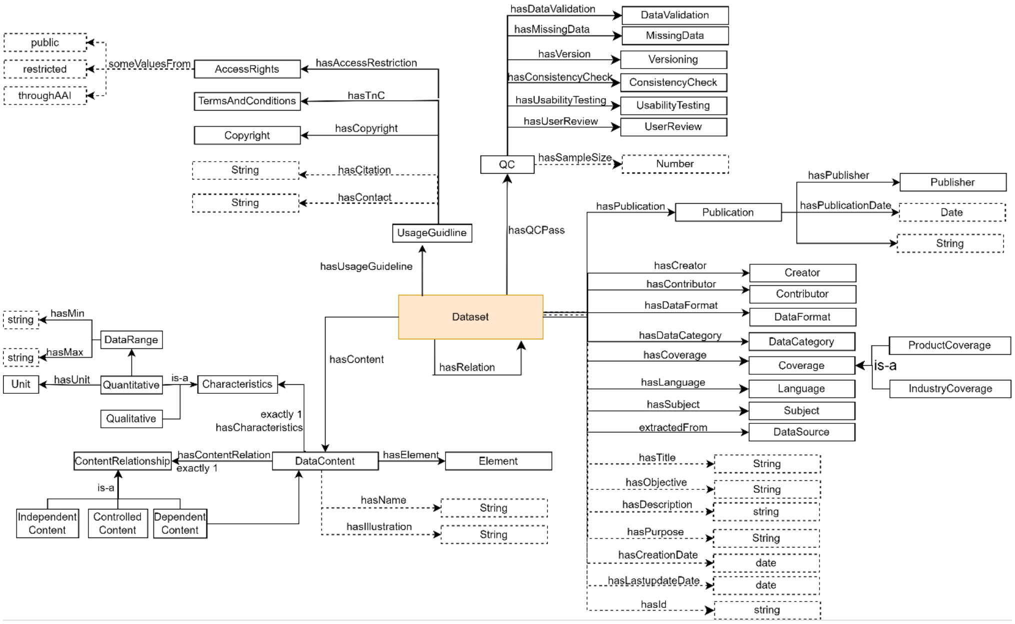


Figure 2. Bird's-eye view of the OntoDataDoc ontology and its facets on data provenance, content and structure, data usage guidelines, and quality check metadata.

The OntoDataDoc ontology focuses on capturing metadata related to data about products, components, and materials, especially composite materials. It adheres to established standards such as Dublin Core and ISO/IEC 11179, incorporating custom concepts and properties to meet specific needs. Different segments of the ontology are elucidated in the following subsections. The ontology defines key classes, including Dataset, Material, Product, Property and Measurement.

5.1.1 Metadata on Data Provenance

The data provenance facet defines several classes, including Creator, Contributor, Publication and Coverage, each serving a distinct role in organising and characterising the data. Object properties, such as hasDataCreator, hasDataContributor, hasMaterial, hasProduct, hasProperty, isPartOf, among others, establish meaningful connections between datasets and their associated elements. The facet also includes datatype properties for representing title, description, objective, date of creation, date of last update, format, subject, identifier, version and source. The facet is depicted in Figure 3, providing an ontological framework for organising, describing, and relating diverse elements within the data documentation.

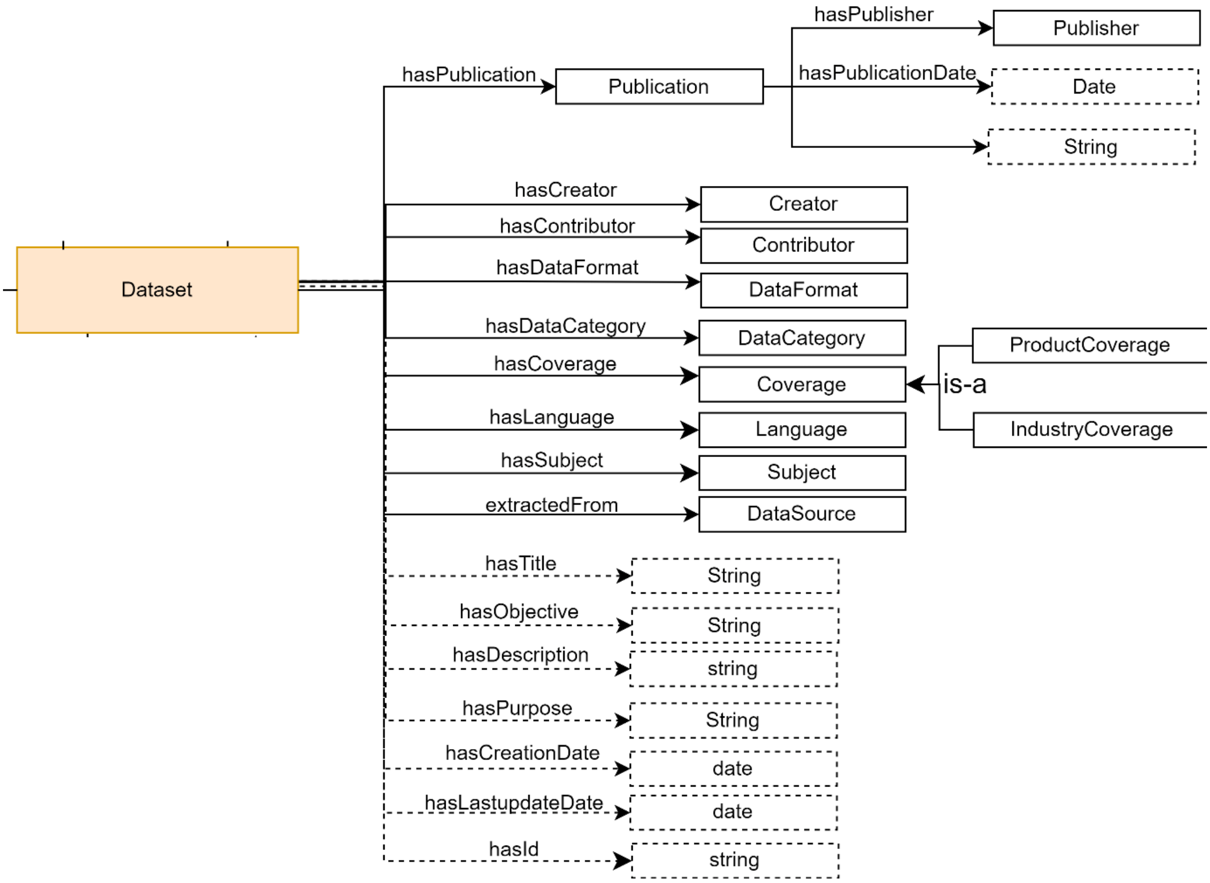


Figure 3. Data Provenance facet of the OntoDataDoc ontology.

5.1.2 Data Content and Structure

The data content and structure facet of the OntoDataDoc ontology is structured around the key class DataContent that is related to the ContentRelationship class that is classified as IndependentContent, The data content and structure facet of the OntoDataDoc ontology is structured around the key class DataContent, which is related to the ContentRelationship class that is classified as IndependentContent, ControlledContent, and DependentContent, as shown in Figure 4. DataContent may have Qualitative or Quantitative characteristics. Quantitative data may have Units and ranges with maximum and minimum values, each playing a specific role in capturing diverse characteristics and properties of products, components and materials. Properties like hasContent, hasValue, hasUnit, hasIllustration, hasName, hasRelation, hasDatatype, hasMin, hasMax, and hasCharacteristics establish meaningful connections and attributes within the ontology. Noteworthy aspects include focusing on variables as central entities and facilitating a comprehensive understanding of materials and products. The ontology supports data integration by linking datasets to their constituent data contents, ensuring effective exploration and analysis. Units and measurement details are captured, promoting consistent and accurate data interpretation. Relationships between pieces of content, data types, value constraints for quantitative contents, and the incorporation of illustrations contribute to the richness of the ontology, enhancing its utility in representing complex concepts and relationships in the realm of product, component and material data description.

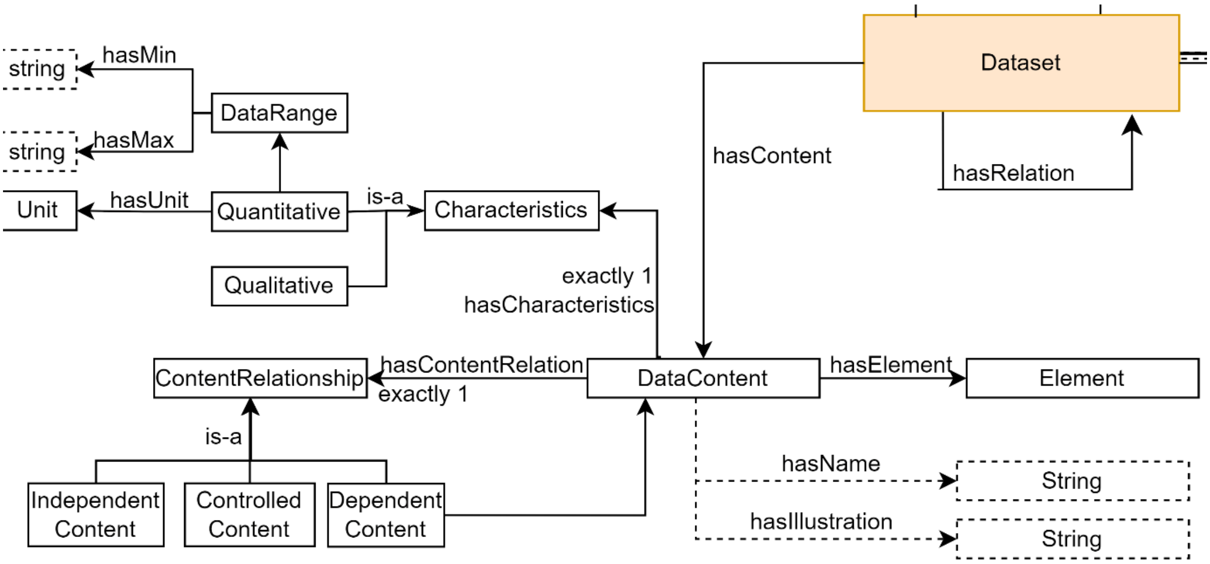


Figure 4. Data Content and Structure facet of the OntoDataDoc ontology.

5.1.3 Quality Check (QC)

The quality check (QC) facet is shown in Figure 5. Quality check methods serve as crucial metadata, ensuring the reliability and usability of data. Data validation involves thorough checks against predefined standards or criteria, such as verifying data points within a specified range or validating correct formatting. Consistency checking ensures coherence within and across datasets, addressing concerns like uniform date formatting or consistent units of measurement. Usability testing assesses the dataset's user-friendliness, examining if users can easily comprehend and utilise the information. User review engages users in identifying errors or issues within the dataset, contributing to accuracy and completeness.

Versioning is employed to track changes made to a dataset over time, enabling the retrieval of previous versions of the dataset if required. Consideration of sample size is integral, as larger sample sizes generally enhance data reliability. Missing data statistics involve monitoring the extent of data absence, a critical aspect for assessing potential impacts on data analysis. These quality check methods can be utilised as metadata, including data validation, consistency checking, usability testing, user review, versioning, sample size, and missing data statistics.

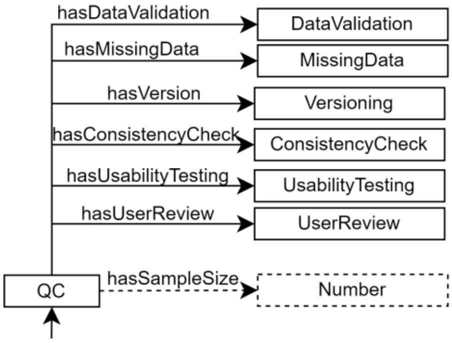


Figure 5. Quality Check facet of the OntoDataDoc ontology.

5.1.4 Usage Guidelines

The usage guidelines facet is portrayed in Figure 6, outlining the core concepts and relationships, along with key aspects of usage guidelines. Core concepts encompass

UsaeGuideline, AccessRight, TermsAndConditions, and Copyright. These concepts form the foundation for effectively managing and documenting data within the ontology. Object properties defined within the facet include hasAccessRestriction, hasTnC, and hasCopyright. Datatype properties include hasCitation and hasContact, describing contact details to enable sending inquiries or seeking permissions of use. The ontology promotes transparency and understanding among stakeholders by providing clear usage guidelines and fostering responsible data-sharing practices with legal compliance. Incorporating usability guidance enhances the ontology's accessibility and usability for data documentation, contributing to effective and ethical data management practices.

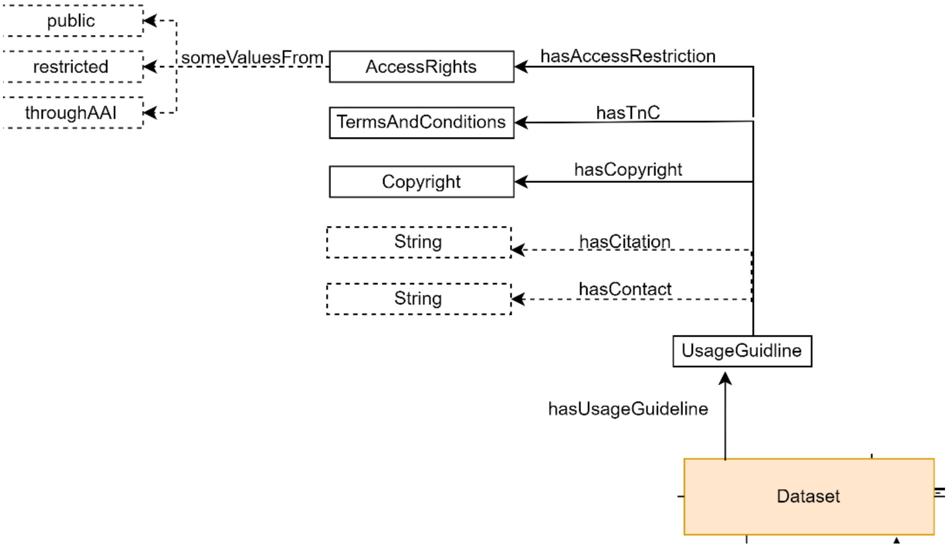


Figure 6. Usage Guidelines facet of the OntoDataDoc ontology.

6. Data Documentation

The data documentation described using Turtle, a variant of OWL and RDF, below showcases how data provenance for JIDEP dataset 1 is represented using the OntoDataDoc ontology. It illustrates how to describe the date of creation and title. The prefix ‘odd’ refers to the IRI of OntoDataDoc.

```
#Data provenance of Dataset 1 of the JIDEP Project
odd:jidep_dataset_1 a odd:Dataset ;
      odd:hasCreationDate “2023-12-31” ;
      odd:hasTitle “Automotive Chassis Cross Beam Dataset” .
```

A dataset usually has different types of contents that may have a complex structure and relationships. The example representation below covers different components, elements, content characteristics, etc.

```
#Data Content
odd:jidep_dataset_1 odd:hasContent odd:crossbeam_01 .
odd:crossbeam_01 a odd:DataContent
      odd:hasName “Chassis Cross Beam” ;
      odd:hasContentRelation odd:cont_rel_03 ;
      odd:hasElement odd:carbon_epoxy ; odd:glass_epoxy .
odd:cont_rel_03 a odd:DependentContent .
```

Usage guidelines may contain metadata to describe a general restriction on usage and who can access the data. Additionally, the terms and conditions are defined from the dataset for guidelines to achieve organisational interoperability. It helps preserve rights such as copyright, appropriate citation requirements, and other legal issues to achieve legal interoperability.

```
#Data Usage Guidelines
odd:jidep_dataset_1 odd:hasUsageGuideline odd:ugl_01 .
odd:ugl_01 a odd:UserGuideline .
      odd:hasAccessRestriction odd:throughAAI ;
      odd:hasTnC odd:prod_terms_conditions_01 ;
      odd:hasCopyright odd:cc4.0 .
```

The turtle snippet below describes data documentation to define data validation, consistency, and testing to build trust and confidence in the dataset.

```
#Data Quality Check
odd:jidep_dataset_1 odd:hasQCPass odd:qc_03 .
odd:qc_03 a odd:QC ;
      odd:hasDataValidation odd:shacl_pass ;
      odd:hasConsistencyCheck odd:quarterly_pass_20240201 ;
      odd:hasUsabilityTesting odd:passed_20240201 ;
      odd:hasSampleSize 5173 .
```

7. Quality Assessment

Quality assessment of the ontology-based data documentation of this project involved two stages: verification and validation. We ensured that the ontology met its specified requirements during the verification process. In this connection, we use OWL reasoners, Pellet [2] and HermiT [3] in Protégé to detect inconsistent assertions. In the validation process, we performed SPARQL on knowledge graphs to validate the quality of the data documentation. The following subsections elaborate on the process of our verification and validation.

7.1 Verification

We created the OntoDataDoc ontology using Protégé, represented it using OWL and applied the following steps for its verification.

1. **Syntax and Structure Verification:**
    - **Visual inspection of syntax and structure**. This ensures the ontology adheres to the OWL syntax rules.
    - **Assessment of naming conventions**. This helps maintain consistency and clarity in the naming of classes and properties.
2. **Semantic Verification**:

    We leveraged SPARQL queries to perform more advanced semantic checks by querying the ontology and verifying specific relationships or the presence of specific datatype properties.

    - **Apply a reasoner to perform consistency checks**. This identifies logical inconsistencies within the ontology, such as contradictory axioms or statements.
3. **Manual Data Instance Verification:**

    We manually reviewed a subset of data instances to ensure their accurate representation and association with the intended classes.

4. Iterative Refinement:

We revised and refined the ontology based on the results of each verification step. This involved fixing syntax errors, adjusting definitions, adding constraints and modifying data instances. We repeated the verification process iteratively until we achieved a satisfactory level of confidence in the correctness, consistency and usability of the ontology.

7.2 Validation

We applied several SPARQL queries to validate the OntoDataDoc ontology and its instantiations. Some of the queries are illustrated below.

This query retrieves datasets, their titles, and copyright information where the rights statement mentions "copyright" (case-insensitive).

```
SELECT ?dataset ?title ?creator ?rights
WHERE {
  ?dataset odd:hasTitle ?title .
  ?dataset odd:hasCopyright ?rights .
FILTER (REGEX(?rights, "copyright", "i"))
}
```

```
SELECT ?dataset ? data_content ?name
WHERE {
  ?dataset odd:hasContent ?data_content .
  ?data_content odd:hasName ?name ;
                odd:hasElement ?element .
}
```

This query retrieves information about data contents that have names and elements.

This query retrieves all contents of datasets where SHACL was used to validate data.

```
SELECT ?dataset ?content
WHERE {
 ?dataset od:hasContent ?content .
 ?dataset od:hasQCPass ?qc .
 ?qc od:hasDataValidation ?consistent
 FILTER (?consistent == "odd:shacl_pass")
}
```

8. Conclusions

The deliverable described OntoDataDoc, an ontology developed for representing data documentation in JIDEP. We extracted required metadata and, reviewed data documentation standards, such as metadata standards, and identified gaps that formed the basis for creating new classes and properties in the ontology. The ontology has several facets, including data provenance, content and structure, usage guidelines, and dataset quality checks. The data provenance focuses on the origin and fundamental information of the data. The data content and structure delineate various variables, specifying their nature, including dependent, independent, and controlled contents. It further encompasses details about datatypes, data ranges, and the nature of the content, whether qualitative or quantitative. In the case of quantitative data, information about associated units of measure is also included. Metadata related to Quality Check (QC) comprehensively outlines data profiling, validation methods, statistical analyses, and various QC tools. Within the Usage Guidelines facet, organisational and legal interoperability properties include policies, terms and conditions, copyrights, citation guidelines, and contact information. The ontology was used to create data documentation of datasets collected from use case partners. The OntoDataDoc ontology will be available at the TheWorldAvatar git repository.

References

[1] H. Rijgersberg, M. Van Assem and J. Top, "Ontology of units of measure and related concepts," *Semantic Web,* vol. 4, no. 1, pp. 3-13, 2013.

[2] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur and Y. Katz, "Pellet: A practical owl-dl reasoner," *Journal of Web Semantics,* vol. 5, no. 2, pp. 51-53, 2007.

[3] B. Glimm, I. Horrocks, B. Motik, G. Stoilos and Z. Wang, "HermiT: an OWL 2 reasoner," *Journal of automated reasoning,* vol. 53, pp. 245-269, 2014.