



MDMC

Master in Data Management
and Curation

MASTER IN DATA MANAGEMENT AND CURATION

Semantic data modeling for interoperable hydrogen research infrastructures

Supervisor(s):
Mariarita DE LUCA,

Candidate:
Smritirekha TALUKDAR

2025–2026





Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA

This course has been funded by the European Union – NextGenerationEU within the projects:

- *NFFA-DI cod. IR00000015 - Missione 4, “Istruzione e Ricerca” – Componente 2, “Dalla ricerca all’impresa” – Linea di investimento 3.1, “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione” – Azione 3.1.1, “Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti” (CUP B53C22004310006)*

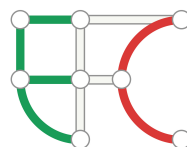
[NFFA-DI : Nano Foundries Fine Analysis - Digital Infrastructure](#)

- *EFC cod. SSU2024 – 00002 - Missione 4, “Istruzione e Ricerca” – Componente 1, “Potenziamento dell’offerta dei servizi all’istruzione: dagli asili nido all’ università” – Linea di investimento 3.4, “Didattica e competenze universitarie avanzate” – Sub-Investimento, “Rafforzamento delle scuole universitarie superiori” (CUP G97G24000100007)*

[EFC : Educating Future Citizen](#)



nffa-di



Author's Declaration

I, SMRITIREKHA TALUKDAR, declare that this thesis entitled, 'Semantic data modeling for interoperable hydrogen research infrastructures' and the work presented therein are my own.

I certify that:

- This work was performed wholly or principally during MDMC internship in THE LABORATORY OF DATA ENGINEERING(LADE), AREA SCIENCE PARK.
- If any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have relied on the published work of others, this is always clearly attributed.
- Where the thesis is based on work I have done in collaboration with others, I have made clear exactly what was done by others and what I contributed.
- With respect to AI technologies:
 - I acknowledge the use of Claude (claude-sonnet-4-6): <https://claude.ai> as a tool to support background research for domain research, all of which were subsequently cross-verified with citeable references and expressed in my own words in the thesis. It was also used to assist me in the grammar and spelling check, paraphrasing of technical content, and suggestions for paragraph organisation throughout this thesis.
 - I acknowledge the use of DeepL (<https://www.deepl.com>) for translation assistance from Italian to English for reference and background documents.
 - No content generated by AI technologies was presented as my own work. All AI-assisted content was critically reviewed, verified against primary sources, and rewritten in my own words before inclusion in this thesis.
- Where I have quoted from the work of others, the source is always cited. Except for such quotations, this thesis is entirely my own work.
- I have acknowledged all major sources of help.

“It was fair enough to write a FAIR thesis. :) ”

Acknowledgments

First and foremost, I would like to thank my supervisor Mariarita de Luca for guiding me throughout the thesis with her valuable guidance and support. I would like to thank Area Science Park for giving me this opportunity to enroll in this course of MDMC which have definitely broadened my knowledge base. The experience and knowledge gained throughout this course will definitely benefit me in my career. Moreover, I would also like to thank Federica Bazzocchi, Stefano Cozzini and all the members of LADE, Area Science Park for their valuable support and help. I also want to thank all the tutors of the course from SISSA, Area Science Park and also from other institutions.

I want to give my sincere gratitude to Alberto Soraci for always being a supportive mentor and Stefano Salvador for the supportive office environment. I also want to thank my colleagues of C1 who has supported me throughout always, especially Francesca(my prior office-mate), Cinzia, Marco, Elisa, Vanessa, Laura, Manuela, Fabrizia, Gabriele, Marinella and the entire office. Moreover, I want to also thank my former supervisor from the University of Trieste, Tiziano Montini for his support throughout my career. Also, grateful for my friends from C11 of UniTs.

I definitely want to thank the people I met in the course who sailed in the same boat with me in the past few months, especially Roberta, Ana, Enrico, Luis, Jacopo, Riccardo, Sandeep. Further, my gratitude towards the warmth, support and love I have always received from my close people in Trieste especially from Manfre and my friends Shahla, Hemanita, Saurabh, Nivetha, Yukti and Devina. Last but not the least I want to thank my parents and my brother and friends from Guwahati for their support in every career step I take.

Abstract

The emergence of hydrogen as an important energy vector has been evidently noted from the past few years. This brings in the importance of research infrastructures dedicated towards hydrogen to further enhance the deployment of this energy vector. These infrastructures, consisting of multiple interconnected units, produce significant amounts of heterogeneous data whose effective utilisation requires a structured and well-defined data management strategy. This work is focused on implementing a FAIR-by-design approach for handling data obtained from a hydrogen smart laboratory which is still under construction and is projected to be completed in 2027. This approach of setting the base and structure for data procurement will save ample amount of time and resources in comparison to future FAIRification of data. Moreover, this approach facilitates for a modular design, thus allowing us to integrate further counterparts and acquire structured data in an elaborated platform. The work consists of three steps following this principle, which involves drafting a DMP, designing a conceptual data model, and then developing a layered, hierarchical ontology based on this conceptual model. The resulting ontology, named HydLab, has been developed in compliance with the FAIR principles, with the additional objective of establishing a modular methodology that enables the extension and adaptation of the model to other hydrogen research infrastructures.

Contents

Introduction	7
1 Context and underlying principles	9
1.1 Context of the thesis	9
1.2 FAIR-data principles	11
1.2.1 What is FAIR-by-design?	12
1.3 Ontologies and their importance in data management	12
1.3.1 Definition and historical background	12
1.3.2 Use of ontologies in Industry 4.0	13
2 Data Management Plan	15
2.1 Introduction to Data Management Plans	15
2.2 Importance of a Data Management Plan in a Research Infrastructure	16
2.3 Methodology for drafting the Data Management Plan	17
2.4 Versioning and future modifications	17
3 Developing the ontology: HydLab	18
3.1 Conceptual data modeling for H_2 research infrastructures	18
3.1.1 Methodology	18
3.2 Design and development of the HydLab ontology	19
3.3 HydLab's compliance with FAIR principles	27
3.4 FAIR-Compliant Dissemination and Repository Organisation	29
4 Conclusions and perspectives	30
A Data Management Plan	33

Introduction

Hydrogen is emerging as an important energy vector in the global transition towards decarbonized energy systems. In an aggregated European level, studies have shown that an integrated hydrogen infrastructure can reduce overall system costs by improving spatial flexibility and facilitating the integration of renewable energy [1]. These research infrastructures can serve as an important link between fundamental science and industrial deployment.

Area Science Park is a public research institution supervised by the Ministry of Universities and Research (MUR). Since 1978, it has managed and developed the Trieste Science and Technology Park, a hub that hosts national and international research centers, innovative companies and startups, and research and technology infrastructure. Recently, it has launched a long-term strategy aimed at creating an advanced hub for research, experimentation, and innovation in hydrogen technologies at the Basovizza Campus. This vision is part of the international framework outlined by the NAHV (North Adriatic Hydrogen Valley) and NASCHA (North Adriatic Smart Communities Hydrogen Accelerator) projects, which contribute to the development of a regional and cross-border ecosystem dedicated to renewable hydrogen and the adoption of innovative technologies in the North Adriatic region. H2SmartLab is a project related to this vision of Area Science Park. It consists mainly of two nodes: the experimental part, which in turn consists of two laboratories: Analysis of Hydrogen Production, Storage, and Use Systems (APSU) and Measurements of the Hydrogen Supply Chain (MFI), both located at the Basovizza campus of Area Science Park. The digital counterpart, Hydor, which will host the computing infrastructure required to create the digital twin and process most of the data acquired, is located at the SISSA – International School for Advanced Studies. The Laboratory of Data Engineering (LADE) of Area Science Park is responsible for the data infrastructure and data management of the project. Integrating data from the node of the physical laboratory to the digital counterpart requires structured, queryable data complying with the FAIR principles. Moreover, a digital twin platform for this lab-scale hydrogen microgrid, will be essential for research,

education, and prototyping of green H_2 microgrid development. This motivation underlies the objective of the thesis, i.e to develop a FAIR-by-design data management and data procurement process for such a physical H_2 microgrid research laboratory. This idea also motivates the overall structure of the thesis.

The thesis is organized as follows:

Chapter 1 gives the context and motivation behind the thesis. This chapter also explains the FAIR-by-design principles and the use of ontologies and their importance in data management. Chapter 2 discusses the Data Management Plan (DMP) for the H2SmartLab and explains the importance of a DMP in a research infrastructure. It also describes the methodology for drafting the DMP and its versioning and future modification details. Chapter 3 explains the development of the ontology *HydLab* and its conceptual data modeling in context to the actual physical laboratory, APSU. The importance of granularity in forming the hierarchical model is also highlighted. This chapter further discusses the choice and importance of re-using existing ontologies for interoperability and the ontology's compliance with the FAIR principles. Chapter 4 contains the conclusions drawn from the results obtained and therefore the future perspectives.

Chapter 1

Context and underlying principles

1.1 Context of the thesis

The [H2SmartLab](https://www.areasciencepark.it/project/h2smartlab/)¹, hydrogen infrastructure system, as described in the introduction, is divided into highly integrated functional components, which contributes to the creation of a unique research and innovation ecosystem: demonstration and experimental work on a real plant and analytical modeling and digital research on the other part.

The experimental component of the infrastructure will be housed in the Technology Center of the Basovizza Campus, where two integrated laboratories will be installed starting in 2026:

- The Analysis, Design, Storage and Use (APSU) laboratory of Area Science Park, which will be dedicated to the experimental study of hydrogen production, storage and use systems using AEM and PEM electrolyzers, hybrid storage systems and fuel cells.
- the Measurements Laboratory for the Hydrogen Supply Chain (MFI) of the University of Trieste consisting of a system of additional and specialized sensors to enhance APSU monitoring through physical and functional measurements along the supply chain.

1. The physical laboratory with all its components will analyze the behavior of the systems under variable operating conditions. This includes critical scenarios, providing the experimental data necessary to evaluate efficiency, operational stability, service continuity and response to technical stress. The construction of this unit is projected to be completed and operational by the end of 2027.

¹<https://www.areasciencepark.it/project/h2smartlab/>

2. The digital component of the physical infrastructure (linked to the APSU and MFI laboratories) is dedicated to the management, curation, analysis, and modeling of experimental data and is essential for transforming the data collected in the laboratories into predictive and operational digital tools to support innovation.

Taking into account the multi-component nature of H2SmartLab, combining physical experimentation, specialized sensing, and digital modeling, there is a clear need for a structured and well-organised data management framework. Since the laboratories are still under development and not yet fully operational, implementing a FAIR-by-design methodology for data acquisition and management enables the integration of semantic interoperability, metadata organization, and monitoring requirements directly into the architectural design of the infrastructure. Embedding these aspects from the beginning, facilitates the future development of interoperable and reusable data ecosystems, while avoiding the structural limitations and costly adaptations often associated with retrospective FAIRification approaches. Digital twins are widely applied to microgrid system analysis with wide applications across wind, solar, hydroelectric, geothermal, and biomass. However, for green hydrogen production and consumption, it still remains emerging. There have been only few reports on green hydrogen and fuel cell microgrid digital twin whereas most of the studies has been on individual components of the physical plant [2] [3] [4]. This brings in future emergence of such more hydrogen research infrastructures incorporating a digital twin microgrid. Moreover, in the context of Industry 4.0, characterized by the integration of technologies such as the Internet of Things (IoT), cloud computing, and advanced automation, the digital twin technology plays a crucial role by establishing a dynamic link between physical assets and their digital counterparts. This aligns very well with the idea behind the H2SmartLab project. Such a dynamic link between the physical and digital units enables detailed analysis and continuous optimization of real-world systems. It ultimately drives improvements and fosters innovation [4].

As mentioned previously, the need for ordered and structured data management requires the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) principles from the point of data acquisition in the physical plant and throughout the subsequent data lifecycle [5]. This approach of managing data with FAIR principles will provide the descriptive and ordered approach for creating the digital twin which will further facilitate the quality and impact of research in this domain. This will be elaborated in the next chapters. Moreover, it will further comply with the reuse of this modality for other such hydrogen research infrastructures. To explain it better, this thesis focuses on setting this basis for managing data from the physical laboratories,

acquired and organized through a scalable, interoperable data management system designed according to FAIR principles. The project H2SmartLab aims in the following. The physical unit present in Basovizza will produce data which as explained before, must be structured, ordered and queryable. These data are to be hosted at the [ORFEO](#)² Data Center where the system will ensure the quality, traceability, and reusability of the data necessary for simulation and optimization activities. The Hydor laboratory at SISSA, in collaboration with the Area's [LADE](#)³ research group, will use this data to build advanced mathematical models, dynamic simulations, and create a digital twin of the experimental plant to replicate the plant's behavior in a virtual environment, analyze operating conditions that are difficult to test in the laboratory, and generate reliable predictions of system performance. This integration will enable virtual stress testing, evaluate complex scenarios, optimize hydrogen production and storage strategies, and support predictive diagnostics, providing a crucial tool for designing resilient and efficient energy infrastructure. The digital component is also projected to be completed and operational by the end of 2027 in accordance with the project proposal.

1.2 FAIR-data principles

A pivotal moment in the development of the FAIR principles came with the 2014 Leiden workshop in the Netherlands, which brought together key stakeholders from academia, industry, and publishing. The workshop underscored the need for a minimal and community-agreed set of guiding principles to reduce the barriers to data discovery, access, and reuse. This culminated in the formal publication of the FAIR principles by Wilkinson et al. in 2016 in *Scientific Data*. In that publication, Mark Wilkinson and colleagues provided the first definitive articulation of Findability, Accessibility, Interoperability, and Reusability as a framework for responsible data stewardship [5]. Thus, this resulted in the formulation of the FAIR principles, which have since been widely endorsed and implemented in various scientific domains. These FAIR principles are further explained on the alignment of the work of this thesis in Chapter 3.

In the following years, the FAIR principles have thus gained considerable momentum in the research world. It can be noted with major funding bodies like the European Commission under Horizon Europe and the National Institutes of Health (NIH), which have progressively embedded FAIR com-

²<https://orfeo-doc.areasciencepark.it/>

³<https://www.areasciencepark.it/en/research-infrastructures/data-engineering-lade/>

pliance into their data management and sharing requirements. This growing policy-level endorsement, coupled with the rapid expansion of data-intensive research fields, has elevated FAIR from a community-agreed guideline to a near-universal standard for scholarly data governance.

1.2.1 What is FAIR-by-design?

FAIR-by-design is a concept implemented in this thesis where the core ambition is to implement techniques that generate FAIR-compliant data directly at the point of acquisition, thereby reducing the time and effort required to re-process existing datasets. This approach further helps us to establish a strong modular foundation which can be expanded following the same conceptual methodology prior to data collection. In this context, the Data Management Plan and the ontology-based semantic model presented in this thesis represent an important step towards embedding FAIR-by-design principles into the data infrastructure of a hydrogen research facility. It would ensure that the data generated across its physical components, from the electrolyzer stacks to the storage unit and balance of plant, is structured, machine-readable, and reusable inherently. This approach is particularly relevant with context to the digital twin model to be built, where the availability of well-structured, interoperable datasets is a definitive prerequisite for building reliable and scalable digital representations of the physical hydrogen plant. The theory of a DMP and its methodology is presented in Chapter 2, however the next section highlights on the definition and importance of ontologies.

1.3 Ontologies and their importance in data management

1.3.1 Definition and historical background

The term "ontology" originates from philosophy which refers to the systematic study of existence and the nature of being. However, its adoption into computer science and information technology (IT) marked a significant conceptual shift. In 1993, Tom Gruber introduced ontology as a technical term in computer science, defining it as a formal description of the concepts and relationships that can exist within a domain of interest. As stated by him in the article, the statement follows as

"A specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects—is called an ontology"[6].

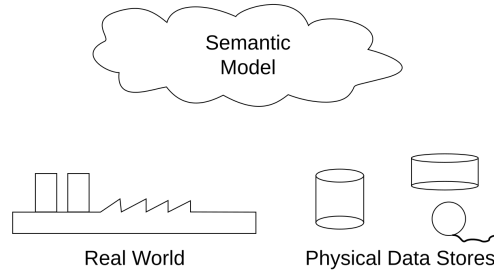


Figure 1.1: Semantic Data Model
[7]

Ontologies are now used to specify standard conceptual vocabularies, which are part of the W3C standards stack for the Semantic Web [8]. These have some key applications such as:

- exchanging data among systems facilitating interoperability
- answering queries
- publishing reusable knowledge bases

These key applications enable the formulation of different semantic data models based on different domains. Moreover, a semantic data model, in terms of software engineering can be explained as a conceptual data model in which semantic information is included.

1.3.2 Use of ontologies in Industry 4.0

Industry 4.0 brings in technological concepts such as Internet of Things (IoT), cloud computing, artificial intelligence, and therefore integrates cyber-physical systems [9]. This leads to the generation of volumes of heterogeneous data across various interconnected devices and platforms. This leads to the need of sharing a common understanding of the structure of information among people or software agents. Thus, this cultivates the primary goal in developing ontologies, enabling reuse of domain knowledge across independently developed systems and services. In the years following Industry 4.0, ontologies have now become a foundational component of the semantic web, Knowledge Graph (KG) engineering, and FAIR data management in scientific research.

In our case, the project H2SmartLab complies as an Industry 4.0 research infrastructure, considering the integration of the physical hydrogen production, usage and storage systems along with the digital unit performing digital

monitoring, control, with data management layers. Moreover, some significant points that make it more compliant are stated below -

- Semantic interoperability in the cyber-physical systems is a vital goal in this project, where ontologies will enable machines, sensors and software from different manufacturers to understand each other. Further, it will enable heterogeneous data from multiple sources and formats to be merged, queried, and reasoned over.
- Knowledge representation of the physical systems will provide the semantic backbone for the digital twin defining what the physical asset is, what properties it has, and how it behaves.
- Standardized vocabularies such as SSN/SOSA, SKOS, QUDT, EMMO, Schema.org etc. enables describing sensors, observations, and actuators in an universal factory environment.
- The FAIR-by-design principle adapted in the work-flow of this thesis directly addresses the Industry 4.0 challenge of managing large volumes of heterogeneous data across interconnected systems.

To align the motivation of incorporating FAIR principles in the data handling of the project, this thesis focuses mainly on

- Drafting a Data Management Plan (DMP)
- Designing a conceptual data model
- Developing a layered, hierarchical ontology : HydLab

The following chapters in the remainder of this thesis is structured to address the three points. The next chapter focuses on the background, theory, methodology of the DMP. It sets the foundation for how data will be collected, stored, described, and shared throughout the project lifecycle.

Chapter 2

Data Management Plan

2.1 Introduction to Data Management Plans

A Data Management Plan (DMP) is a structured, formal document that describes how data will be handled throughout the lifecycle of a research project and even beyond. It formalizes the proper collection, organization, storage, and preservation of the research data. It is a document far from being a mere bureaucratic "box-ticking" exercise attached to funded projects, and rather is a methodological best practice applicable to all research activities [10]. In recent years, DMPs have gained a great deal of importance due to the growing emphasis on research transparency, reproducibility, and open science [11]. Many funding agencies and institutions now require researchers to submit a DMP as part of their proposals. Moreover, A DMP supports compliance with ethical and legal standards while also improving the reliability and long-term value of research outputs by establishing clear procedures for handling data.

Further, DMPs have become mandatory for all Horizon Europe projects at present, unlike its predecessor Horizon 2020, where an opt-out option existed. Similarly, the European Research Council (ERC) has mandated applicants to submit a Research Data Management plan as of 2021. The Austrian Science Fund (FWF) has required a data management plan for all funded projects since 2019, in which researchers describe how data are collected, organised, stored, published, shared, and archived. More examples across Europe include the Netherlands Organisation for Scientific Research (NWO), which mandates open FAIR data practices, the German Research Foundation (DFG), which publishes explicit guidelines on the handling of research data, the Swiss National Science Foundation (SNSF), which requires open research data planning, and the Research Council of Norway (RCN),

which mandates open access to research data. This convergence, particularly across European funding bodies, reflects a shared commitment to open science principles and signals that DMPs have become a baseline expectation of any funding body [12] [13] [14] [15] [16] [17] [18].

2.2 Importance of a Data Management Plan in a Research Infrastructure

A well structured and comprehensive DMP is a very critical component of any modern research infrastructure. Designing a DMP around the FAIR principles, implements the idea of FAIR-by-design methodology which embeds the principles directly into the research workflow, thereby avoiding the resource-intensive process of retroactive FAIRification of data. In this regard, a comprehensive DMP serves several fundamental purposes:

- **Data Description and Collection:** It clearly defines the types, formats, and volumes of data to be generated or collected, establishing a shared understanding and details of the data involved in the research infrastructure.
- **Metadata and Standards:** It specifies the metadata schemas and controlled vocabularies to be adopted, ensuring that data remains interpretable and self-describing across independently developed systems and services.
- **Data Access and Privacy:** It clearly defines who can access the data, under what conditions, and through which protocols, thereby safeguarding both the integrity of the research and the rights of all personnels and partners involved.
- **Storage, Backup and Preservation:** It addresses where data will be stored, how it will be backed up, and for how long it will be retained, ensuring continuity and resilience of the research data infrastructure.
- **Legal and Ethical Compliance:** It accounts for intellectual property rights, licensing conditions, and any applicable data protection regulations, ensuring that the management of data remains compliant with institutional and legal frameworks.
- **Roles and Responsibilities:** It assigns clear ownership and accountability for data management tasks across the different personnels and

partners involved, from data producers to repository managers and institutional head.

Beyond its initial formulation, the periodic revision and updating of the DMP is equally essential. As the laboratory evolves in terms of its technological infrastructure, research scope, and institutional partnerships, the DMP must be treated as a living document, ensuring that data management practices remain aligned with the growing complexity of the laboratory's operations and the ever-advancing standards of open and responsible science.

2.3 Methodology for drafting the Data Management Plan

The first draft of the DMP was drafted with a structured and consultative process. The methodology involved a systematic review of relevant official documents such as the project proposal of H2SmartLab, Piping and Instrumentation Diagram (P&ID) and technical mappings of the laboratory, institutional guidelines to give a foundational understanding of the regulatory and operational requirements. Along with that, active participation in meetings with the engineering consultants of the laboratory allowed for the incorporation of technical perspectives and domain-specific constraints. Further discussion with the senior laboratory management ensured that the DMP aligns with the institution's strategic priorities, while remaining grounded in the operational realities of its research infrastructure.

2.4 Versioning and future modifications

The DMP for the project, H2SmartLab is provided in the Appendix of this thesis. It is also available in the GitHub repository of the project. Since this project is ongoing and will undergo more developments, the first draft was created following the methodology described above. This DMP as stated before is a living document and will again undergo future revisions and will have further modifications.

The following chapter explains the next steps for the FAIR-by-design approach by explaining the designing of the conceptual model which follows towards the development of a layered, hierarchical ontology, HydLab.

Chapter 3

Developing the ontology: HydLab

In the previous chapter, we had encountered the first step towards FAIR-by-design approach, i.e. drafting a data management plan. As explained before, this chapter lets us go through the proceeding steps of designing a conceptual data model and developing a layered, hierarchical ontology for such hydrogen research infrastructures.

3.1 Conceptual data modeling for H_2 research infrastructures

Conceptual data models are basically graphical or textual data models designed for data modeling at a high level of data abstraction. These data models help to capture, elaborate and describe rich and complex semantics [19]. When designed, keeping in mind the FAIR-by-design approach to be followed, it ensures that the semantic relationships governing the data are formally assigned and captured at the earliest stage of the data lifecycle.

3.1.1 Methodology

The development of the conceptual data model (CDM) presented in this thesis followed a structured step-by-step methodology. The steps involved were as follows -

- In the context of the hydrogen laboratory in this work, the initial step was going through the Piping and Instrumentation Diagram (P&ID) of the physical plant under construction. Each physical and functional component of the laboratory was identified with domain knowledge. Further, a flowchart, figure 3.1 was followed to decide the placement

of each of these nodes, which determined the hierarchical position of each entity based on its functional role and its dependencies with respect to other components. Prior to the flowchart, the six layers were defined from the methodological point of view. The flowchart also decided whether each particular node was a class or an instance (named individual).

- This structured approach, consistently applied to the different components of the physical laboratory, enables the scalable integration of additional elements within the hierarchical structure, making the model adaptable to other similar research infrastructures. Furthermore, it contributed to defining the boundary constraints and institutional scope of the data models.
- Following the flowchart, six hierarchical levels were declared. These six levels have been represented in the form of a Knowledge Graph (KG), which accurately reflects the operational logic of the hydrogen research infrastructure, rather than just its mere physical topology. It also provides a detailed analogy of each node realised through mapping which further extends the generalisation of the KG, as it is shown in figure 3.2.
- Construction of this KG required the application of domain knowledge along with exploring the granularity of each level where future addition of more components to this model makes it entirely possible. This means that this class structure is in a layered manner with assigned hierarchy which makes this model both reusable and interoperable for other hydrogen research infrastructures. The multi-layered semantic structure of HydLab follows a modular design strategy intended to facilitate the scalable integration of additional components and the adaptation of the ontology to other hydrogen research infrastructures and related energy systems.

3.2 Design and development of the HydLab ontology

The conversion from the conceptual knowledge graph to a formal ontological representation constituted the next methodological step. This included the mapping of the semantic structure of the hydrogen research infrastructure onto existing and newly defined ontological frameworks. Primary step of

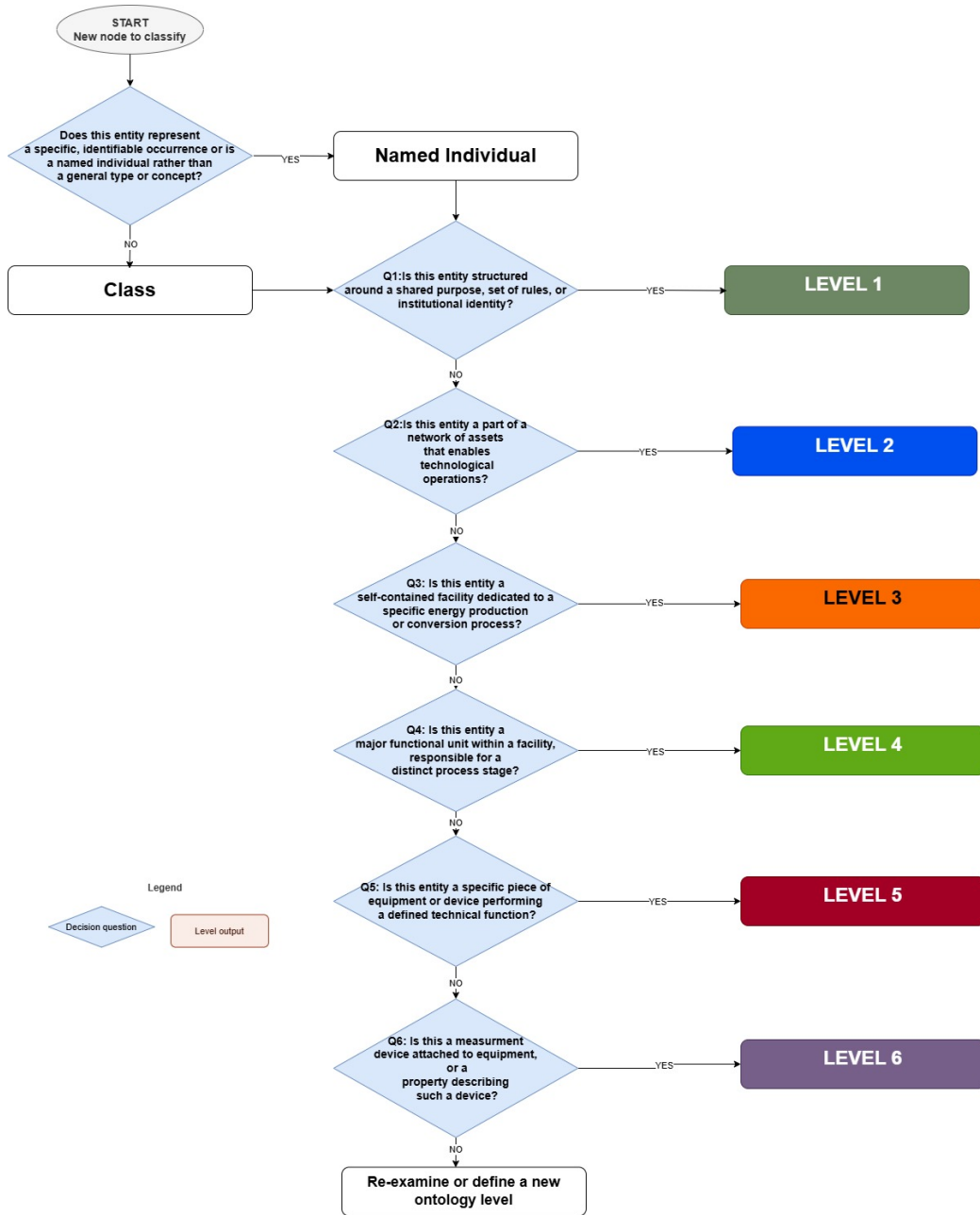


Figure 3.1: Flowchart for node entry into levels. The different levels are color coded according to hierarchy as are to be represented in the KG.

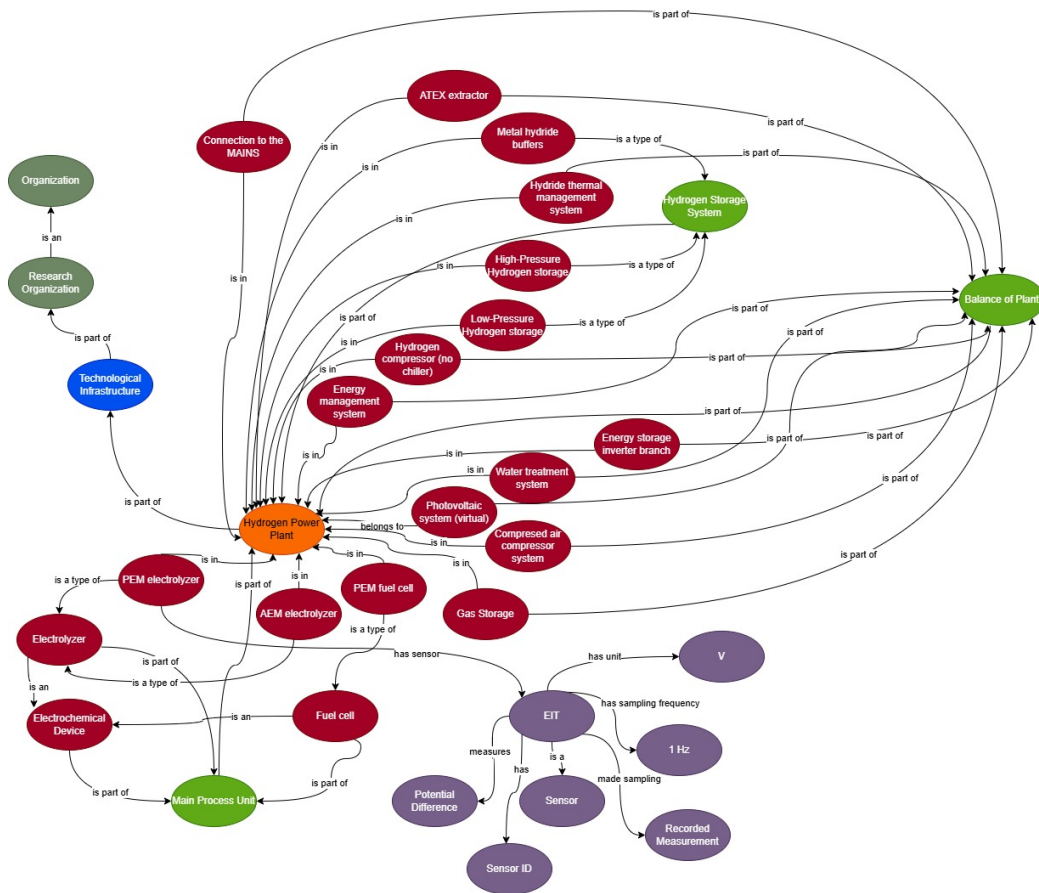


Figure 3.2: Knowledge Graph of a Hydrogen Research Infrastructure. The colors in this KG are linked to the different levels as shown in figure 3.1

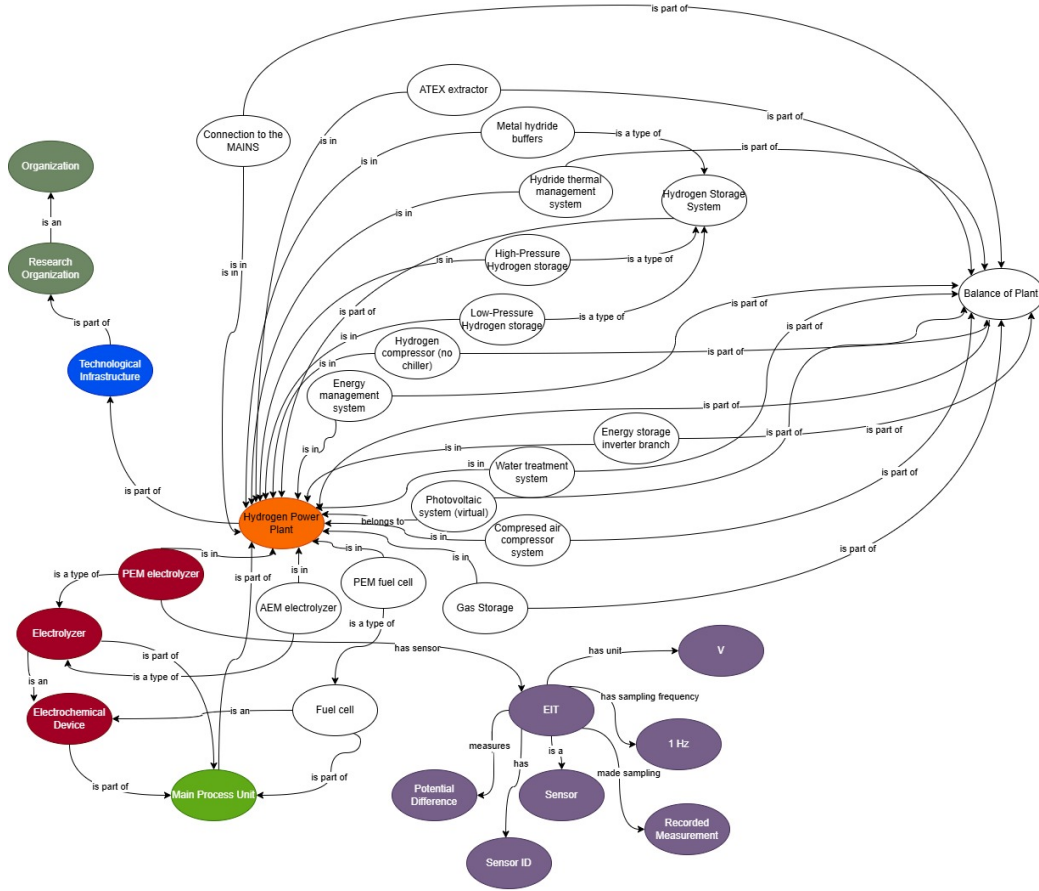


Figure 3.3: One branch highlighted of the KG

this was to focus on one branch of the conceptual model as highlighted in the figure 3.3.

The focus on this particular branch with explored granularity led to a process of ontological mapping that was sourced from a multi-source in depth research process:

- **Linked Open Data (LOD) resources:** The LOD ecosystem was consulted to identify existing ontologies and vocabularies relevant to the domain. It was also ensured that the resulting ontological structure adheres to open and interoperable standards.
- **Existing ontologies and their documentation:** Relevant ontologies were examined through their associated scientific publications. It provided both the theoretical grounding and the semantic scope necessary to assess their applicability to the hydrogen research domain.

- GitHub repositories: The development versions and technical specifications of candidate ontologies were studied directly from their respective repositories, which led to a detailed assessment of their class structures, object properties, and namespace declarations.

The narrowing of the scope focusing on just one single branch allowed for the comprehensive definition of semantic relationships between the nodes explaining one specific sub-domain. This logic can be further extended to the other branches of the entire KG thus leading to the generalization of this approach. This will ensure methodological consistency across the entire ontology.

To explain the ontology diagram shown in Figure 3.4, it can be seen that the six hierarchical levels have been color coded. These levels had been classified for each node following the flowchart of 3.1. The Level 1 and the outmost layer undergoes the classification for institution, underneath which follows the next layer of level 2 classification i.e. the subdivision of other technical infrastructures that could be present under the same organization. Level 3 constitutes an energy production or conversion facility which in our case is the hydrogen power plant, APSU. The level 4 classifies as the units present in the energy plant which are responsible for a specific process stage. Proceeding towards level 5, it constitutes a specific piece of equipment or device which is a part of the level 4 unit. Level 6 constitutes a specific measurement device, which in the branch explained in this work is the Electrical Impedance Tomography (EIT), along with the properties describing the device. Each of these levels are connected through the conceptual realizations in the KG which are then assigned with some existing ontologies.

The existing ontologies re-used in the ontology diagram shown in figure 3.4 are explained below. Further, these have been grouped according to their functional role within the overall knowledge representation framework.

- W3C Core Standards : These provide the logical, syntactic, and schematic primitives necessary for any OWL-compliant ontology
 - OWL (Web Ontology Language): It constitutes the formal backbone of the ontology, providing the logical constructors necessary for defining classes, object properties, datatype properties, and named individuals [20].
 - RDF (Resource Description Framework): It provides the foundational triple-based data model upon which the entire ontology is serialized. It gives the representation of all entities and their interrelations as subject-predicate-object statements [21].

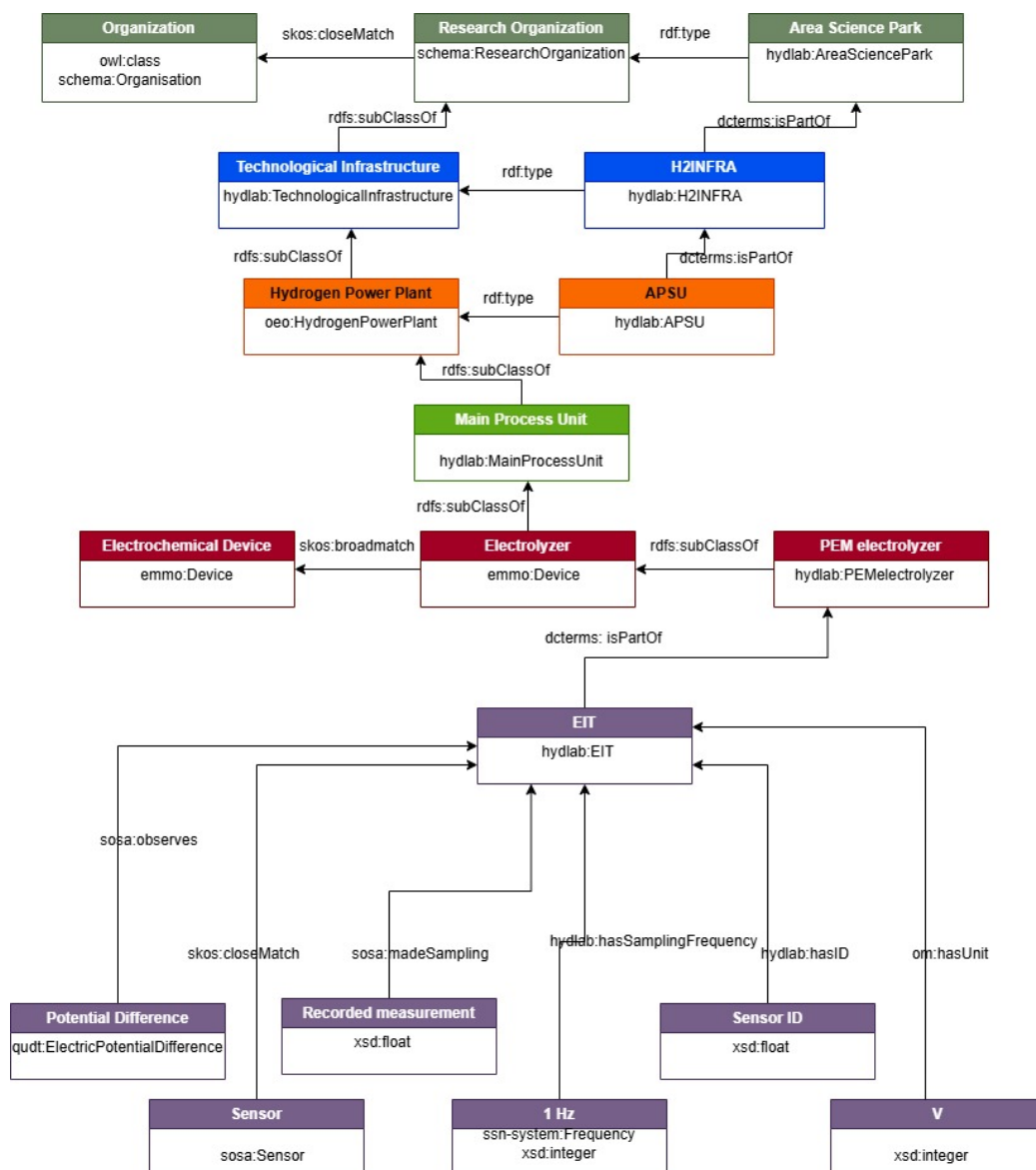


Figure 3.4: Ontology Diagram

- RDFS (RDF Schema): It extends RDF by providing the vocabulary necessary to define class hierarchies and property domains and ranges [22].
 - XSD (XML Schema Definition): It supplies the datatype vocabulary used to formally type literal values throughout the ontology, including numerical measurements, identifiers, and date values [23].
 - SKOS (Simple Knowledge Organization System): It supports conceptual alignment and labelling across the ontology, enabling the association of classes with broader concepts, alternative labels, and human-readable definitions [24].
 - SOSA (Sensor, Observation, Sample and Actuator): Provides the semantic framework for representing the sensing and observation layer of the hydrogen laboratory in the case of our work [25].
 - SSN (Semantic Sensor Network): It complements SOSA by providing an extended vocabulary for describing the capabilities and properties of sensing systems [26].
- Metadata and findability vocabularies : These ontologies govern the descriptive and bibliographic metadata of the ontology itself.
 - DCTerms (Dublin Core Metadata Terms): It provides the metadata vocabulary used to annotate the ontology, capturing descriptive information such as title, description, creation date, and licensing. More notably, `dcterms:isPartOf` is reused as the primary structural relation in our case of forming the hierarchical KG [27].
 - VANN (Vocabulary for Annotating Vocabulary Descriptions): Provides the metadata properties used to declare the preferred namespace prefix and URI of the ontology [28].
 - Domain ontologies : These ontologies provide domain-specific semantic alignment, grounding the classes and properties of `hydlab` within established vocabularies relevant to the hydrogen, energy, materials science, and organisational domains.
 - Schema.org: It serves as the reference vocabulary for the organisational layer of the ontology [29].
 - EMMO (Elementary Multiperspective Material Ontology): It provides the device-level ontological alignment through `emmo:Device`,

for an electrochemical device in our work, within a materials science and engineering perspective that can be adopted across other research infrastructures with ease [30].

- OEO (Open Energy Ontology): It provides the domain-specific class `oeo:HydrogenPowerPlant`, enabling the alignment of the hydrogen laboratory’s power plant component with an open energy domain [31].
- Measurement and Units Ontologies: These ontologies ensure that all physical quantities and their associated units of measure are formally and unambiguously represented in accordance with scientific standards.
 - OM-2 (Ontology of Units of Measure): It provides the property `om:hasUnit` used to formally associate measurement quantities with their respective units of measure [32].
 - QUDT (Quantities, Units, Dimensions and Types): It supplies the formal class definition for physical quantities. In our work, `qudt:ElectricPotentialDifference` is used to semantically ground the observable property measured by the electrochemical impedance tool within a widely adopted scientific measurement vocabulary [33].

The choice of the aforementioned ontologies was realized to have been directly governed by the hierarchical structure of the conceptual knowledge graph. This highlights the granularity importance of the KG since from the organizational context at the uppermost level down to the sensor measurement apparatus at the lowest, required layers of semantic alignment.

Therefore, a dedicated namespace was declared which served as the primary semantic space within which the ontology, HydLab was developed. This namespace acts as the foundational identifier for all classes, object properties, and individuals defined within the ontology. It however includes and aligns the external ontologies through established ontology integration practices such as import statements and property mappings. Within the domain of hydrogen energy, which is an upcoming important energy vector, HOLY exists which focuses mainly on the the structure, actors, products, applications, and market dynamics of the hydrogen economy to support structured analysis, integration, and discovery of hydrogen market data [34]. Thus, an ontology for hydrogen research infrastructure was necessary to be developed for making the data in this domain more semantically rich and FAIR compliant.

Below is a section dedicated to explaining how HydLab follows the FAIR principles.

3.3 HydLab's compliance with FAIR principles

In this section, we describe in detail the adherence of HydLab to the FAIR principles and the strategies adopted to support findability, accessibility, interoperability, and reusability within the proposed semantic framework.

- To be Findable:
 - **F1** - It states that the (meta)data must be assigned a globally unique and persistent identifier. The ontology, HydLab will be assigned a persistent identifier through its Zenodo record, which generates a DOI upon upload. This DOI serves as the globally unique and persistent identifier for both the ontology and its associated metadata. The versioning and update of the content in the GitHub repository doesn't affect the DOI and the Zenodo record.
 - **F2** - The data must be described with rich metadata (defined by R1 below). Here, rich metadata will be provided through the Zenodo record, including title, description, creators, keywords, license, and version information.
 - **F3** - The metadata clearly and explicitly must include the identifier of the data it describes. In our case the ontology file itself, points to its permanency and citation via the DOI of the Zenodo record. Moreover, any metadata record for describing the ontology, has the namespace URI as the related identifier.
 - **F4** - The (meta)data are registered or indexed in a searchable resource. By publishing on Zenodo and linking the GitHub repository, the ontology becomes indexed in a searchable open repository, further enhanced by the LOD ecosystem through the declared namespace.
- To be Accessible:
 - **A1** - The (meta)data are retrievable by their identifier using a standardized communications protocol. The ontology is retrievable via the standard protocol used by zenodo and GitHub.
 - **A1.1** - The protocol must be open, free, and universally implementable. Both of the Zenodo DOI and the GitHub URL operate over HTTPS, which is an open, standardised, and universally implemented communication protocol.
 - **A1.2** - The protocol must allow for an authentication and authorization procedure, where necessary. The ontology is openly

accessible through Zenodo and GitHub, however, these platforms allow the possibility to support access control mechanisms where necessary.

- **A2** - The metadata must be accessible, even when the data are no longer available. Even if the ontology file itself is updated or deprecated, the metadata deposited on Zenodo remains accessible.
- To be Interoperable:
 - **I1** - The (meta)data must use a formal, accessible, shared, and broadly applicable language for knowledge representation. The ontology is serialised in OWL/Turtle which is a broadly adopted language for knowledge representation, compliant with W3C standards.
 - **I2** - The (meta)data must use vocabularies that follow FAIR principles. The ontology reuses vocabularies that are themselves FAIR-compliant, including OWL, RDF, RDFS, SKOS, SOSA, SSN, DCTerms, and Schema.org, all of which are maintained under open and persistent URIs.
 - **I3** - The (meta)data include qualified references to other (meta)data. Through the systematic reuse of and alignment with external ontologies such as EMMO, OEO, QUDT, and OM-2, the ontology includes qualified references to other data and metadata resources.
- To be Reusable:
 - **R1** - The meta(data) must be richly described with a plurality of accurate and relevant attributes. The ontology is richly described both internally through OWL annotations and externally through Zenodo metadata.
 - **R1.1** - The (meta)data must be released with a clear and accessible data usage license. The ontology will be released under an open license declared both within the ontology file and on the Zenodo record.
 - **R1.2** - The (meta)data must be associated with detailed provenance. Provenance information is captured through DCTerms properties within the ontology itself, including creation date, authorship, and the institutional context of its development, ensuring full traceability of the ontological artefact.

- **R1.3** - The (meta)data must meet domain-relevant community standards. The ontology, HydLab, follows domain-relevant community standards by aligning with established ontologies in the hydrogen, energy, sensor, and materials science domains. Moreover, the vocabularies that have been reused have been done so with appropriate licenses declared thoroughly in the GitHub repository.

3.4 FAIR-Compliant Dissemination and Repository Organisation

The research results produced in the context of this thesis are organized across two dedicated and interlinked GitHub repositories, each serving a distinct purpose. The primary thesis repository contains the Data Management Plan drafted for the hydrogen laboratory which is a part of the project H2SmartLab, alongside the corresponding README and license documentation. Rather than duplicating the ontological artefacts, the thesis repository maintains a direct reference to the dedicated HydLab ontology repository, which constitutes the canonical source for all ontology-related files. The HydLab repository is structured such that the Turtle serialization of the ontology (ontology/hydlab.ttl) serves as the single source of truth and is only editable, from which the OWL/RDF XML and JSON-LD representations are automatically derived through a Python conversion script. Human-readable HTML documentation is further generated and served through the docs/index.html file, providing an accessible entry point for both domain experts and developers. The GitHub Actions continuous integration workflow ensures that the ontology is automatically validated and all serialisation formats are regenerated upon every push to the repository, maintaining consistency across all representations at all times. Both repositories are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license and are archived on Zenodo with persistent digital object identifiers, ensuring their long-term findability, accessibility, and citability in full compliance with the FAIR principles. The link for the GitHub repository of the thesis is [MDMC-thesis-2025-26-Talukdar GitHub repository](#)

Chapter 4

Conclusions and perspectives

The motivation towards an ordered approach for structuring and managing data for hydrogen research infrastructures, thus resulting into a dynamic link between physical assets and their digital counterparts was highlighted in the introduction of this thesis. The development of HydLab followed this same motivation with a methodologically grounded approach that reflects some established practices in ontology engineering and semantic web research. Listed below are some conclusions derived from the workflow of this thesis -

- The entire thesis was motivated by FAIR-by-design methodology which can be reflected starting from the Data Management Plan (DMP) which has been drafted prior to the existence of a physical laboratory. The drafting of the DMP, carried out through continuous interaction with the project manager, participation in engineering meetings, and analysis of the official project documentation, contributed to establishing a robust foundation for FAIR data management by embedding data organization, semantic interoperability, and metadata requirements directly into the infrastructure planning process. The ontology development also followed the same as the basis started from constructing a conceptual knowledge graph (KG). This knowledge graph (KG) was inspired and derived from the Piping and Instrumentation Diagram (P&ID) of the hydrogen plant, related to the project of H2SmartLab, currently being developed at the institution. This ensured that the semantic structure of the ontology derived from this KG, faithfully reflected the physical and functional architecture of an actual hydrogen research infrastructure.
- The ontology development of HydLab followed some gold-standard practices such as implementing the NeOn methodology and the LOD

(Linked Open Data) principles. More iteratively, the NeOn methodology or Network of Ontologies for ontology engineering highlights nine scenarios for ontology development, which in the case of HydLab closely takes into account Scenarios 2 and 5 [35] [36]. Scenario 2 suggests the reuse and re-engineering of existing ontologies whereas scenario 5 suggests the reuse and merging of existing ontologies. In the case of HydLab, reused and partially re-engineered existing ontologies (SOSA, SSN, EMMO, OEO, Dublin Core) were merged into a new domain-specific ontology for hydrogen research infrastructures.

- The ontology model is kept open for modification and further versioning. This is not only for the future modifications of the physical laboratory to be constructed in the institution but also for other hydrogen research facilities. This reflects the METHONTOLOGY approach which is well established in the semantic field. This methodology comments that the ontology life cycle should be based on evolving prototypes [37].

The work carried out in this thesis sets the semantic and data management foundation upon which a fully interoperable and FAIR-compliant hydrogen research infrastructure can progressively be foreseen. This will further facilitate the building of the digital twin at the Hydor node for the H2SmartLab project. An important aspect of developing HydLab is that it establishes the semantic foundation necessary for linking the physical hydrogen research infrastructure to its future digital twin. Moreover, an immediate objective of the ontology that has been built involves the generation of a JSON-LD serialization for the semantic enrichment of the datasets generated by the Programmable Logic Controller (PLC) of the physical laboratory. This will enable the datasets to be not only machine-readable but also semantically grounded, thus allowing us to query and retrieve information through SPARQL, which is a standardised query language for RDF-based knowledge graphs. Therefore, the raw laboratory measurements will be queryable as well as machine interpretable. Another future perspective of this work is to validate the logical consistency of the ontology through automated reasoning via Pellet or HermiT in Protégé to verify no logical contradictions. This is to ensure that HydLab remains a contradiction-free and extensible knowledge representation framework as the physical infrastructure evolves. Moreover, beyond the institutional context, thanks to its modular and layered structure, HydLab's ontological framework can be potentially broadly adapted within the hydrogen research community. This approach promotes semantic harmonization across diverse energy research infrastructures, fostering interoperability with emerging digital twin frameworks and federated data spaces developed to support the interconnection of data and research

infrastructures at broader international scale, such as the European Open Science Cloud (EOSC).

Appendix A

Data Management Plan

This Data Management Plan (DMP), encoded as DMP_H2SmartLab 1.0 is the first draft for the H2SmartLab project.

Data description and collection or re-use of existing data

a) How will new data be collected or produced and/or how will existing data be re-used?

The new data will be produced from the APSU (for the analysis of hydrogen production, storage and use systems) and MFI (for additional measurements on the hydrogen supply chain) units which is the first node called the H2 Integrated Lab of the workflow-pipeline. The data ingestion process will be managed through the Head PLC operating in read-only mode, ensuring that the data acquisition infrastructure cannot directly interact with or modify the plant automation system for safety and operational reliability reasons. A modular and automated data pipeline will collect the measurements generated by the APSU and MFI instrumentation and transfer them to a Time Series Database (TSDB) hosted within the ORFEO datacentre for storage, monitoring, and further semantic enrichment processes.

The re-use of the existing data could be done by extracting from the data collector, saved in the format of a time-series database which will be enriched with metadata and then further stored in a datacentre ORFEO (database to be decided).

b) What data (for example the kind, formats, and volumes) will be collected or produced?

The data will be collected in CSV format exported by the central PLC of the APSU and MFI instrumentation. The sampling frequency of the

data will be 1 Hz per sensor for 98 sensors and 2 Hz for 5 sensors. The ratio of data writing performance to machine load has to be optimised.

The methodologies or software that will be used if new data are collected or produced are explained below. The method for data provenance documentation is also explained.

The operational data pipeline is intended to rely on a Kafka-based streaming architecture designed to manage real-time data events generated by the laboratory instrumentation. The pipeline will follow a modular structure, enabling the separation of data ingestion, data cleaning, metadata enrichment, and storage functionalities. During the enrichment phase, the acquired measurements will be associated with a metadata schema derived from the semantic metadata model defined through the HydLab ontology. The enriched metadata and semantic mappings will be represented through JSON-LD serializations to support interoperability, machine-readability, and semantic integration across heterogeneous systems. The semantically enriched data streams will then be stored within a Time Series Database (TSDB) hosted in the ORFEO datacentre, supporting scalable storage, monitoring, interoperability, and future digital twin integration workflows. The proposed data flow is designed to support future interaction and synchronization mechanisms with the digital twin environment.

The ontological mapping relation with RDF triples can be also pictorially represented (.png)

The details of the volumes is explained below.

The volume of the data collected will depend on the number of sensors placed and the frequency with which these sensors are collecting the data.

The data volume in case the collection frequency is 1 Hz for a sensor can be calculated as -

$$\text{Data volume (bytes/s)} = \text{Sample size (bytes)} \times \text{Sampling rate (Hz)} \quad (\text{A.1})$$

The number of sensors along with their sampling rate is as such:

$$98 \text{ sensors} \rightarrow 1 \text{ Hz} \quad (\text{A.2})$$

$$5 \text{ sensors} \rightarrow 2 \text{ Hz} \quad (\text{A.3})$$

Hence,

$$\text{Data volume (/s)} = \text{Sample size (bytes)} \times \{(98 \times 1) + (5 \times 2)\} \text{ Hz} \quad (\text{A.4})$$

$$= \text{Sample size (bytes)} \times 108 \text{ Hz} \quad (\text{A.5})$$

For the 5 sensors collecting data in Boolean,

$$\text{Data volume (/s)} = 10 \text{ samples/s} \times 1 \text{ byte} \quad (\text{A.6})$$

$$= 10 \text{ bytes/s} \quad (\text{A.7})$$

For the other 98 numeric sensors collecting data in REAL/INT/FLOAT
 Minimum estimate:

$$\text{Data volume (/s)} = (98 \text{ samples/s} \times 4 \text{ bytes}) + 10 \text{ bytes/s} \quad (\text{A.8})$$

$$= 402 \text{ bytes/s} \quad (\text{A.9})$$

Maximum estimate (if stored as double):

$$\text{Data volume (/s)} = (98 \times 8 \text{ bytes}) + 10 \text{ bytes/s} \quad (\text{A.10})$$

$$= 794 \text{ bytes/s} \quad (\text{A.11})$$

This volume of data is an estimated calculation before the metadata enrichment. This data volume is estimated from the central PLC.

Documentation and data quality

a) What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data

The metadata will be linked to the variables of the parameters measured by the sensors through IRIs. This will enable the data to be queryable. The documentation regarding the methodology of the data collection, organisation of the data and the scripts developed for converting CSV to JSON-LD will be stored in the GitHub repository of the project named as H2SmartLab. The documentation for data schema and ontology mapping will also be presented in the GitHub repository of the project.

b) What data quality control measures will be used?

Data quality checks will be integrated directly within the operational pipeline through modular validation stages designed to identify inconsistencies, missing values, duplicated timestamps, anomalous measurements, and communication errors during the streaming process. The implementation of these quality-control procedures follows principles inspired by industrial data quality pipelines proposed in recent literature for real-time industrial environments [38]. The validation and cleaning stages will therefore operate as an integral component of the streaming architecture before metadata enrichment and long-term storage within the ORFEO infrastructure.

Storage and backup during the research process**a) How will data and metadata be stored and backed up during the research process?**

The raw and semantically enriched data streams will be stored within a Time Series Database (TSDB) hosted in the ORFEO datacentre, which will act as the primary storage infrastructure for both operational and curated data generated by the APSU and MFI units. Temporary buffering and streaming management may be implemented within the ingestion pipeline to support reliable real-time acquisition and processing workflows before long-term storage.

At the current stage of the project, both the physical laboratory infrastructure and the final production-ready database environment are still under development. Consequently, the definitive backup, replication, and disaster recovery strategies have not yet been fully established and will be defined during the commissioning and operational deployment phases of the infrastructure. These strategies will be developed in coordination with the ORFEO infrastructure management team, taking into account scalability requirements, long-term preservation needs, data integrity, and operational reliability of the H2SmartLab ecosystem.

b) How will data security and protection of sensitive data be taken care of during the research?

The data-security will be maintained with API keys between the data flow in the APSU, MFI and the Hydor Unit of the Hydrogen plant. Personal data management and data security procedures will comply with Regulation (EU) 2016/679 (GDPR), together with applicable institutional and national data protection policies.

The legal and ethical requirements, together with the related codes of conduct governing data management and access procedures within the H2SmartLab ecosystem, will be formally defined through the future H2SmartLab data policy, which is currently under development. The policy will regulate aspects related to data ownership, access permissions, intellectual property rights, data sharing conditions, and the management of potentially sensitive operational information.

The codes of conduct adopted within the project will be aligned with the institutional regulations and ethical guidelines applicable to the participating organisations. In particular, research activities and data management practices will follow the principles defined in the codes of conduct for researchers and in the internal code of conduct governing the personnel of Area Science Park, together with any additional institutional and collaborative agreements established among the project partners.

Data sharing and long-term preservation

a) How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

The conditions for sharing of the data is to be agreed upon the three partners of the project, i.e. AREA Science Park, SISSA and University of Trieste. The project aims to make documentation, metadata, and selected datasets openly accessible whenever possible, while access to operational and raw experimental data may be regulated according to the project data policy and access control procedures. The documentation and the metadata enrichment project will be available to be accessed openly in the GitHub repository of the project. The project follows a FAIR-by-design approach.

b) How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

At the current stage of the project, the long-term preservation strategy and the criteria for selecting data for permanent archival have not yet been fully defined, as both the physical infrastructure and the operational data ecosystem are still under development. The present workflow foresees that raw data generated by the main PLC will be streamed through a Kafka-based pipeline for data validation, cleaning, and semantic metadata enrichment before being stored within a Time Series Database (TSDB) hosted in the ORFEO datacentre.

The long-term preservation policies, including the selection criteria for raw, curated, and semantically enriched datasets, as well as the identifi-

cation of potential repositories or archival infrastructures, will be established during the commissioning and operational phases of H2SmartLab in coordination with the project partners and the ORFEO infrastructure management team. The preservation strategy will take into account scalability, FAIR compliance, interoperability requirements, long-term scientific reuse, and future digital twin integration needs.

c) What methods or software tools will be needed to access and use the data?

The data stored within the Time Series Database (TSDB) hosted in the ORFEO datacentre will be accessible through query interfaces compatible with SQL or SQL-like languages, depending on the final database technology adopted within the infrastructure. Access to semantically enriched datasets and metadata may additionally rely on APIs and semantic interoperability services developed within the H2SmartLab ecosystem.

The TSDB environment will implement a Role-Based Access Control (RBAC) system to ensure controlled and differentiated access to the various data layers according to user roles, responsibilities, and authorization levels. The authentication and authorization mechanisms are expected to be integrated with the ORFEO Authentication and Authorization Infrastructure (AAI), enabling centralized identity management and secure access procedures across the infrastructure.

d) How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

During the operational phases of the project, the data streams and records stored within the Time Series Database (TSDB) will be associated with unique internal identifiers to ensure traceability, consistency, and linkage between raw measurements, metadata, and semantically enriched representations across the data pipeline.

For datasets selected for dissemination, publication, or long-term preservation, persistent identifiers such as Digital Object Identifiers (DOIs) will be assigned through suitable trusted repositories or archival platforms adopted by the project partners. The published datasets will be indexed and described through administrative and descriptive metadata including:

project initiation date and dataset description; author(s) and contributors; funding information; proposal or project identifiers, where applica-

ble; persistent identifiers associated with publications or dissemination activities; researcher identifiers (e.g. ORCID).

Datasets made openly available will be distributed under the CC-BY (Creative Commons Attribution) license or other equivalent open-access licenses compatible with the project data policy and applicable institutional regulations.

Data management responsibilities and resources

a) Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)

The first draft of the Data Management Plan (DMP) has been developed within the framework of the Master in Data Management and Curation (MDMC) by a researcher affiliated with Area Science Park and involved in the H2SmartLab activities. The present DMP primarily addresses the APSU and MFI units managed by Area Science Park and the University of Trieste, while future complementary DMP sections related to the Hydor unit will be developed in collaboration with SISSA.

The DMP and the associated data-management workflows will be progressively reviewed and refined in coordination with the project manager and the responsible personnel of the APSU, MFI, and Hydor units as the infrastructure evolves toward operational deployment.

The operational data management activities will subsequently be supervised by dedicated data stewardship roles responsible for monitoring the organization of the data ecosystem, including documentation, metadata management, version control, semantic interoperability, storage organization, and FAIR compliance procedures. Since the physical infrastructure and the associated digital ecosystem are still under development, the data-management procedures and governance mechanisms will be periodically reviewed and updated during the commissioning and operational phases of the project.

b) What resources (for example financial and time) will be dedicated to data management

At the current stage of the project, the first implementation activities related to data management, semantic enrichment, and metadata modeling are being developed within the framework of the MDMC internship activities. These activities include the de-

velopment of preliminary data pipelines, metadata schemas, ontology mappings, semantic enrichment workflows, and FAIR-oriented documentation procedures.

The operational pipeline is expected to rely on a modular Kafka-based streaming architecture designed for real-time data ingestion, validation, cleaning, metadata enrichment, and storage. Dedicated Python scripts and automated services will process the raw data generated by the PLC systems and transform them into semantically enriched representations based on metadata schemas derived from the HydLab ontology framework.

To support FAIR compliance, the project foresees the adoption of the following measures:

- Findable - The datasets and metadata will be associated with unique identifiers and enriched with ontology-based semantic metadata, including IRIs, provenance information, timestamps, sensor identifiers, and standardized units of measurement. Published datasets may additionally receive persistent identifiers such as DOIs through trusted repositories.
- Accessible - The data ecosystem will rely on the ORFEO data-centre infrastructure and on controlled access mechanisms based on Role-Based Access Control (RBAC) integrated with the ORFEO Authentication and Authorization Infrastructure (AAI). Access conditions and permissions will be regulated according to the future H2SmartLab data policy.
- Interoperable - Interoperability will be supported through the adoption of widely used ontologies and semantic standards, including Schema.org, HOLY, QUDT, UCUM, SOSA/SSN, SensorML, EMMO, BFO, and FOAF. Metadata enrichment will rely on JSON-LD serializations and ontology mappings derived from the HydLab semantic framework, enabling semantic integration across heterogeneous infrastructures and future interoperability with broader federated research ecosystems.
- Reusable - The data-management architecture is expected to follow a **Medallion-style layered organization** within the ORFEO infrastructure:
 - the Bronze layer will contain raw acquisition data exported from the instrumentation together with basic provenance metadata;

- the Silver layer will contain cleaned, validated, and standardized time-series data enriched with harmonized meta-data and semantic identifiers;
- the Gold layer will contain semantically enriched and interoperable datasets represented through JSON-LD and RDF-based semantic structures suitable for advanced querying, analytics, and future digital twin integration workflows.

The final allocation of long-term financial, infrastructural, and personnel resources dedicated to operational data stewardship, storage, preservation, and FAIR compliance will be defined during the future operational phases of the H2SmartLab infrastructure.

Bibliography

- [1] J. Radek, M. S. Breder, and C. Weber, “Hydrogen in the european power sector – a case study on the impacts of regulatory frameworks for green hydrogen,” *Energy Policy* **213** (6, 2026) 115200. <https://linkinghub.elsevier.com/retrieve/pii/S0301421526001345>.
- [2] T. Clennar, Y. Du, and Z. Liu, “Lab-scale green hydrogen and fuel cell microgrid digital twin,” *Energy Conversion and Management* **350** (2, 2026) 120928. <https://www.sciencedirect.com/science/article/pii/S0196890425014529>.
- [3] D. Calderón, F. J. Folgado, I. González, and A. J. Calderón, “Development and real-world implementation of a digital twin-enhanced supervisory system for res-green hydrogen microgrid,” *Energy Conversion and Management* **348** (1, 2026) 120822. <https://www.sciencedirect.com/science/article/pii/S0196890425013469>.
- [4] Z. Feng, I. Eiubovi, Y. Shao, Z. Fan, and R. Tan, “Review of digital twin technology applications in hydrogen energy,” *CHAIN* **1** (7, 2024) 54–74.
- [5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. V. D. Lei, E. V. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg,

- K. Wolstencroft, J. Zhao, and B. Mons, “Comment: The fair guiding principles for scientific data management and stewardship,” *Scientific Data* **3** (3, 2016) .
- [6] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowledge Acquisition* **5** (6, 1993) 199–220.
<https://www.sciencedirect.com/science/article/pii/S1042814383710083>.
- [7] N. I. of Standards Computer Systems Laboratory and Technology, “Fips publication 184: Integration definition for information modeling (idefix),” tech. rep., National Institute of Standards and Technology (NIST), 12, 1993.
<https://web.archive.org/web/20131203223034/http://www.itl.nist.gov/fipspubs/idefix.doc>. Archived at the Wayback Machine on 2013-12-03. Originally published by the Computer Systems Laboratory of NIST.
- [8] T. Gruber, *Ontology*. Springer, 2016.
<https://tomgruber.org/writing/definition-of-ontology/>. Accessed: 2026-05-04.
- [9] K. Schwab, *The Fourth Industrial Revolution*. Random House LLC, 2016.
- [10] U. degli Studi di Firenze, “Data management plan (dmp),” 3, 2026. Open Science section, Research and Innovation. Last modified: March 2026. Accessed: 2026-05-08.
- [11] K. Briney, H. Coates, and A. Goben, “Foundational practices of research data management,” *Research Ideas and Outcomes* **6** (7, 2020) .
- [12] R. C. of Norway (RCN), “Open access to research data,” 2023.
<https://www.forskningsradet.no/en/research-policy-strategy/open-science/research-data/>. Accessed: 2026-05-08.
- [13] S. N. S. F. (SNSF), “Open research data: Data management plan,” 2017. <https://www.snf.ch/en/dMILj9t4LNk8NwyR/topic/open-research-data>. Accessed: 2026-05-08.
- [14] D. F. (DFG), “Guidelines on the handling of research data,” 2015. <https://www.dfg.de/en/research-funding/funding-initiative/research-data/activity-areas>. Accessed: 2026-05-08.

- [15] N. O. for Scientific Research (NWO), “Research data management policy,” 2016.
<https://www.nwo.nl/en/research-data-management>.
Accessed: 2026-05-08.
- [16] A. S. F. (FWF), “Research data management at the fwf,” 2019.
<https://www.fwf.ac.at/en/research-funding/open-access-policy/research-data-management>. Accessed: 2026-05-08.
- [17] E. R. Council, “Open research data and data management plans: Information for erc grantees,” 2021.
https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf. Accessed: 2026-05-08.
- [18] E. Commission, “Horizon europe programme guide: Open science and data management,” 2021. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf. Accessed: 2026-05-08.
- [19] Z. Ma and L. Yan, “Data modeling and querying with fuzzy sets: A systematic survey,” *Fuzzy Sets and Systems* **445** (9, 2022) 147–183. <https://www.sciencedirect.com/science/article/pii/S0165011422000203>.
- [20] World Wide Web Consortium (W3C), “Owl 2 web ontology language document overview,” 2012.
<https://www.w3.org/TR/owl2-overview/>. Accessed: 2026-05-08.
- [21] World Wide Web Consortium (W3C), “Rdf 1.1 concepts and abstract syntax,” 2014.
<https://www.w3.org/TR/rdf11-concepts/>. Accessed: 2026-05-08.
- [22] World Wide Web Consortium (W3C), “Rdf schema 1.1,” 2014.
<https://www.w3.org/TR/rdf-schema/>. Accessed: 2026-05-08.
- [23] World Wide Web Consortium (W3C), “Xml schema definition language (xsd) 1.1,” 2012.
<https://www.w3.org/TR/xmlschema11-1/>. Accessed: 2026-05-08.
- [24] World Wide Web Consortium (W3C), “Skos simple knowledge organization system reference,” 2009.

- <https://www.w3.org/TR/skos-reference/>. Accessed: 2026-05-08.
- [25] World Wide Web Consortium (W3C), “Semantic sensor network ontology — sosa,” 2017. <https://www.w3.org/TR/vocab-ssn/>. Accessed: 2026-05-08.
- [26] World Wide Web Consortium (W3C), “Semantic sensor network ontology — ssn,” 2017. <https://www.w3.org/TR/vocab-ssn/>. Accessed: 2026-05-08.
- [27] Dublin Core Metadata Initiative (DCMI), “Dcml metadata terms,” 2020. <https://www.dublincore.org/specifications/dublin-core/dcml-terms/>. Accessed: 2026-05-08.
- [28] I. Davis, “Vann: A vocabulary for annotating vocabulary descriptions,” 2010. <https://vocab.org/vann/>. Accessed: 2026-05-08.
- [29] Schema.org Community Group, “Schema.org — schemas,” 2024. <https://schema.org>. Accessed: 2026-05-08.
- [30] European Materials Modelling Council (EMMC), “Elementary multiperspective material ontology (emmo),” 2023. <https://emmo.info/emmo>. Accessed: 2026-05-08.
- [31] Open Energy Family, “Open energy ontology (oeo),” 2021. <https://openenergy-platform.org/ontology/oeo/>. Accessed: 2026-05-08.
- [32] H. Rijgersberg, M. van Assem, and J. Top, “Ontology of units of measure (om 2),” 2017. <http://www.ontology-of-units-of-measure.org/resource/om-2/>. Accessed: 2026-05-08.
- [33] QUDT.org, “Qudt — quantities, units, dimensions and types,” 2023. <https://qudt.org>. Accessed: 2026-05-08.
- [34] K. M. A. Arevalo, C. Neunsinger, R. Zimmermann, R. Blum, and K. Weakly, “Holy: An ontology covering the hydrogen market,” in *The Semantic Web – ISWC 2023, Lecture Notes in Computer Science*, pp. 3–17. Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47243-5_1.
- [35] M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, *Ontology engineering in a networked world*. Springer Berlin Heidelberg, 1, 2012.

- [36] T. Berners-Lee, “Linked data,” 2006.
<https://www.w3.org/DesignIssues/LinkedData.html>. Design
Issues, W3C.
- [37] M. Fernández-López and A. Gómez-Pérez, “Overview and
analysis of methodologies for building ontologies,” 2002.
- [38] T. Peixoto, O. Oliveira, E. Costa e Silva, B. Oliveira, and
F. Ribeiro, “A data quality pipeline for industrial environments:
Architecture and implementation,” *Computers* **14** no. 7, (June,
2025) 241. <http://dx.doi.org/10.3390/computers14070241>.