D3.1 – 1st release of the DIGITISE Data Space Implementation

Project number	101160671
Project acronym:	DIGITISE
Project title:	Digital Innovative cross-sector services for Greater citizen Integration in a just energy Transltion, and Societal Empowerment

Work Package	WP3: Data Management, Interoperability, Sharing and AI Baseline Analytics
Responsible Partner	Suite5
Official Submission Date	31/08/2025
Actual Submission Date	01/09/2025
Type	Other (O)
Dissemination Level	Public (PU)
Reviewers	UNP, UBI
Version	V1.0

Funded by the European Union

Versioning and contribution history

Version	Date	Author(s)	Notes
v0.1	12/2024	Suite5	ToC
v0.2	3/2025	Suite5, Tech partners	Standards review
v0.3	4/2025	Demo partners	Data needs
v0.4	5/2025	Suite5, Tech partners	DIGITISE data model draft version
v0.5	5/2025	Suite5, Tech partners	Components ToC
v0.6	6/2025	Suite5, Tech partners	Components 1st version
v0.7	31/7/2025	Suite5, Tech partners	Final version for review (data model and components)
v0.8	15/8/2025	UNP, CIRCE	Review
v0.9	29/8/2025	Suite5	Address comments
v1.0	1/9/2025	UBI	Submission

Preface

Funded by the European Commission under Grant Agreement number 101160671, DIGITISE is a project focused on enhancing the digital literacy and empowerment of consumers and prosumers in the energy sector. By integrating advanced technologies and fostering active engagement in digital energy activities and markets, DIGITISE aims to play a crucial role in the global energy transition.

Executive Summary

This deliverable reports on the activities and progress within WP3 of the DIGITISE project as of M15. The central objective of WP3 is to design and implement a robust, interoperable, and intelligent data ecosystem that serves as the foundational data infrastructure for the entire project, supporting seamless data exchange and advanced analytics across multiple sectors.

Key achievements detailed in this document span the full data lifecycle. The work began with the establishment of a comprehensive DIGITISE Data Model (Task 3.1), which ensures semantic interoperability by adapting open standards from initiatives and the standardization (IEC, OPENADR etc..) and modeling new concepts specific to the project's needs. The overall work performed takes into account also the data needs of the project. At the end, the 3 domain models are defined for the different sectors of interest in DIGITISE. Following the definition of the models, the methodology and tools for the lifecycle management of the different sectorial models is provided.

To populate the data space, dedicated data connectors and robust governance mechanisms have been implemented (Task 3.2), enabling the secure ingestion of data from diverse sources, including physical assets and third-party APIs. The data collection process is supported by semantic harmonization and data curation techniques while further results are stored to a dedicated storage layer (Along with the associated metadata).

The scope of the work in Task 3.3 as implemented also for the data space environment is twofold: to facilitate data sharing and commerce, a Data Marketplace has been created (Task 3.3), featuring a search and exploration engine, smart contract management, and secure API-based data retrieval governed. Moreover, to enhance privacy and security over the data advanced Attribute-based Access Control (ABAC) and personal data sovereignty rules are applied further accompanied by a novel data anonymization service.

The different services mentioned above are complemented by the DevOps environment that will ensure the prompt deployment of the data space to mandate the execution of the different services as well as monitoring the prompt performance of the overall system. The entire ecosystem is managed and coordinated by a Data Space Operations Layer (Task 3.5), which orchestrates resources, monitors performance, and ensures the stable, secure, and scalable operation of all components.

Following the DIGITISE data space specifications, a significant innovation within WP3 is the development of AI-enabled analytics pipelines (Task 3.4), which provide both personal and energy-specific insights, such as demand forecasting and flexibility profiling. A list of analytics models needed in the project has been delivered and an early primitive version of the base models has been made available in this first version. Moreover, and with focus on the integration activities to be delivered in the project, the dedicated framework for the integration is provided as part of the work in the early phase.

Funded by the European Union

In summary, this document presents the architecture, implementation details, and operational considerations of the foundational data space, marking a significant step toward achieving the project's ambitious goals for cross-sectoral digital transformation. This work is pivotal for the DIGITISE project, as it provides the curated, interoperable, and governed data assets (and analytics services) that are essential for the development of advanced services and business applications in subsequent work packages (WP4 and WP5).

Funded by
the European Union

Table of Contents

Funded by the European Union

List of Figures

List of Tables

Funded by the European Union

Abbreviations

Abbreviation	Full Name
ABAC	Attribute-based Access Control
AI	Artificial Intelligence
AMI	Advanced Metering Infrastructure
API	Application Programming Interface
APM	Access Policy Management
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BEMS	Building Energy Management Systems
BIM	Building Information Modeling
BOT	Building Topology Ontology
BSD	Berkeley Software Distribution (License)
CAPEX	Capital Expenditures
CEMS	Customer Energy Management Systems
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CIM	Common Information Model
CNNs	Convolutional Neural Networks
COBie	Construction-Operations Building information exchange
CPs	Charging Points
CSA	Connectivity Standards Alliance
DER	Distributed Energy Resources
DevOps	Development and Operations
DHW	Domestic Hot Water
DiCon	Digital Construction Ontologies
DNAS	Drivers, Needs, Actions, Systems
DoA	Description of Action
DR	Demand Response
DRL	Deep Reinforcement Learning
DTs	Digital Twins
EDMC	Enterprise Data Management Council
EIR	Employer's Information Requirements
EMQX	(A specific MQTT Broker, proper name)
ESCO	Energy Service Company
ETSI	European Telecommunications Standards Institute
EU	European Union
EV	Electric Vehicle
EXI	Efficient XML Interchange
FBC	Finance Business & Commerce (FIBO Module)
FIBO	Financial Industry Business Ontology
FND	Foundations (FIBO Module)

GBM	Gradient Boosting Machine
H2020	Horizon 2020 (EU Framework Programme)
HVAC	Heating, Ventilation, and Air Conditioning
I/O	Input/Output
IAQ	Indoor Air Quality
IDSA	International Data Spaces Association
IEA	International Energy Agency
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IFC	Industry Foundation Classes
IFRS	International Financial Reporting Standards
IoT	Internet of Things
IRR	Internal Rate of Return
ISO	International Organization for Standardization
IT	Information Technology
JTC	Joint Technical Committee
JWT	JSON Web Token
KEDA	Kubernetes-based Event Driven Autoscaling
KPIs	Key Performance Indicators
LCOE	Levelized Cost of Energy
LEC	Local Energy Communities
LGPL	(GNU) Lesser General Public License
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MQTT	Message Queuing Telemetry Transport
NPV	Net Present Value
NREL	National Renewable Energy Laboratory
obXML	occupant behavior XML
OCPP	Open Charge Point Protocol
OPC UA	Open Platform Communications Unified Architecture
OpenADR	Open Automated Demand Response
OPEX	Operating Expenses
PINN	Physics-Informed Neural Network
PM	Particulate Matter
PPAs	Power Purchase Agreements
PubSub	Publish-Subscribe
PV	Photovoltaic
REG	Regulatory & Compliance (FIBO Module)
RES	Renewable Energy Sources
RMSE	Root Mean Square Error
SAREF	Smart Applications REFerence ontology
SEC	Securities (FIBO Module)

SGAM	Smart Grid Architecture Model
SSO	Single Sign-On
TOU	Time Of Use
TVOC	Total Volatile Organic Compounds
V2G	Vehicle-to-Grid
VOCs	Volatile Organic Compounds
VPP	Virtual Power Plant
WP	Work Package
XBRL	eXtensible Business Reporting Language
XML	eXtensible Markup Language

1 Introduction

In this deliverable we reported the activities in M15 for Work Package 3 (WP3) of the DIGITISE project dedicated to establishing a robust, interoperable, and intelligent data ecosystem that supports seamless data exchange and advanced analytics across multiple sectors (Tasks 3.1–3.5).

The work begins with a thorough examination of the data assets and requirements necessary to realize the DIGITISE vision (Task 3.1). This involves identifying key concepts and attributes to be integrated into the DIGITISE Data Model, ensuring semantic compatibility among diverse data sources and stakeholders (Task 3.1). By studying data modeling practices in energy, health, finance, and other sectors, and leveraging initiatives from the International Data Spaces Association (IDSA[1]) and GAIA-X[2], the project selects and adapts open standards, semantic models, and ontologies[3][4] to meet its baseline needs (Task 3.1). At the same time, the team addresses the need to model new concepts—such as flexibility markets, storage, and security data—that may not be fully represented in existing models (Task 3.1).

Parallel to this, WP3 develops dedicated connectors and data exchange protocols to enable formalized, interoperable ingestion of data from physical assets at the consumer side, as well as from external sources such as open data repositories (Task 3.2). Building on the big data platform developed in the H2020-SYNERGY [5] project, new data ingestion features are integrated to ensure compliance with open standards (such as Modbus [6]for distributed energy resources and OCPP [7] for electric vehicle chargers) (Task 3.2). The project also implements robust data governance mechanisms to manage data collected in various forms, whether through batch uploads, third-party APIs, or real-time streaming via PubSub mechanisms, ensuring effective and secure data handling within the DIGITISE data platform (Task 3.2).

To facilitate the discovery, sharing, and controlled access to data assets, WP3 delivers a Data Search & Exploration Engine that allows users to efficiently search for and visually explore data within the DIGITISE environment, always respecting the access policies associated with each asset (Task 3.3). This capability is embedded within a Data Marketplace, which not only manages the contractual aspects of data sharing—offering smart contract templates and legally binding terms—but also encourages active consumer participation as data providers (Task 3.3). Secure data retrieval services are implemented to allow self-service access to data via custom and open APIs, all governed by advanced access control mechanisms (Task 3.3). Attribute-based Access Control (ABAC) [8] is introduced as a primary method for enforcing access policies, while personal data sovereignty features empower consumers to define their own anonymization and privacy rules, safeguarding their data against unintended disclosure (Task 3.3).

Funded by the European Union

A key innovation in WP3 is the development of AI-enabled analytics pipelines that bring intelligence to the data space (Task 3.4). These AI components provide both personal analytics—addressing consumer energy behavior, comfort preferences, and occupancy profiling—and energy analytics, such as demand and generation forecasting, flexibility profiling, and context-aware analysis (Task 3.4). The goal is to establish realistic and objective baselines for energy exchanges and flexibility

Finally, WP3 ensures the continuous integration and evolution of all developed components within the DIGITISE data space (Task 3.5). The Data Space Operations Layer is established to manage and coordinate the various sub-components and services, orchestrate resources, and maintain secure, scalable operation. This task also supports the ongoing integration of solutions developed in subsequent work packages, while providing intuitive visualization tools for monitoring and assessing the performance of the entire DIGITISE ecosystem (Task 3.5). While at the early stage of integration actions, some details about the approach to be followed are reported in this deliverable.

In summary, WP3 (Tasks 3.1–3.5) lays the groundwork for a dynamic, interoperable, and intelligent data environment that is essential for enabling the cross-sectoral digital transformation envisioned by the DIGITISE project. The development activities performed to support this data governance and management framework are reported in details in this version of the document.

1.1 Relevance to other Deliverables

This deliverable is related to many activities and work packages in the DIGITISE project. This facilitates the development of the data space, in accordance with the project, to ensure deliverables meet stakeholder needs, implementation is consistent with prior decision making, and development remains on track technically. The activities informing WP3 are from the output from WP2, and, in particular, functional and non-functional requirements applied in Task 2.4, to inform the design and implementation of the data space. This will ensure that the artifactual architecture of the data space is relevant to the market-wide and pilot-specific business decisions, and considers any regulatory considerations from WP2, particularly around data governance and market rules. In addition to the direct input from WP2, WP3 deliverable is also intended to provide a major enabler for the technology and business work packages to follow. The data space engineered in WP3 will provide curated, interoperable and governed data assets to WP4 and WP5, enabling advanced analysis, as well as AI-based and business applications. This will ensure the technical solutions and services developed in WP4 and WP5 will be built upon standards-based data infrastructure, in the future, enable seamless integration and consistent trajectory towards the deliverables for the DIGITISE project.

1.2 Structure of the Document

This deliverable is divided into the following main sections

- Section 1 will be the introduction that will establish the context for this deliverable. It will identify the purpose and objectives of the deliverable and its relevance to the DIGITISE project.
- Section 2 will be the State-of-the-Art analysis, where we will discuss Sectorial Data Models Handling and Interoperability Management; data landscape mapping and tactical overview of the DIGITISE data model and understanding of the approaches to Semantic Interoperability.
- Section 3 will present Data Collection and Governance and the architecture, functions, technical details, and consideration/constraints that must be taken into account for the data collection, data harmonization, data curation and data storage modules.
- Section 4 will address Data Privacy and Sovereignty. It will review the mechanisms and elements for ensuring data privacy, data access and availability, and identity management of data sharing within the DIGITISE data space.
- Section 5 will examine Information Sharing and the Data Marketplace Environment; architecture and features covering the data marketplace, data exploration and data retrieval modules, including implementation and operational considerations.
- Section 6 will outline analytics models/big-data models developed within DIGITISE, detailing the models/tool development, implementation and integration model to the data space.
- Section 7 will explore DIGITISE Integrated Framework covering the integration and operational model for the DIGITISE data space, covering execution management, resource management, operations monitoring.
- Section 8, the main findings and recommendations for future work will be presented.

This approach ensures that each aspect of the DIGITISE data space is addressed and clearly presented in the document.

2 Sectorial Data Models Handling and Interoperability Management

Task 3.1 work focuses on creating an interconnected robust data model in this section. DIGITISE requires the establishment of a comprehensive data model which serves as its foundational element. The process ensures that every piece of project information regardless of its source understands the same language framework. The primary goal focuses on semantic compatibility to make sure that all project components including various data sources and stakeholders interpret data identically. This method serves as an essential tool to stop data misunderstandings while enabling smooth information transfer.

The project does not seek to create entirely new frameworks but remains dedicated to exploring current data modeling practices found in energy, healthcare and finance sectors. The project will use initiatives from the International Data Spaces Association (IDSA) to determine and adapt existing open standards, semantic models and ontologies. The project uses a strategic method to achieve its fundamental data requirements in an efficient and effective way.

Furthermore, acknowledging the dynamic evolution of the energy landscape, the project must also develop models for novel concepts. This includes critical areas such as flexibility markets, energy storage solutions, and robust security data, which may not be comprehensively represented in existing models. The analysis will take into consideration the review of existing data assets in the project in order to address the needs of the DIGITISE framework.

Through the execution of these steps, DIGITISE is aimed to create a data model that is not only thorough but also fully interoperable to address the project needs and beyond.

2.1 State of the art analysis

Starting with the state-of-the-art analysis, the focus is on the review of the relevant standardization on the domains of interest in the project. The main innovation of the approach is the focus on other than energy domains, namely the financial and health/security as of interest of the project. The analysis per sector is presented below.

2.1.1 Energy domain standardization overview

In this section, we aim to address the energy related needs of the project by reviewing the relevant standardization work. Modern energy systems increasingly rely on digital assets—heterogeneous devices operating in real-time. This mass penetration of assets creates semantic interoperability challenges, as conflicting data models and standards hinder seamless integration. Existing approaches, such as the SGAM model developed under CEN-CENELEC/ETSI Mandate M490 and the IEC Common Information Model (combining

IEC 62325, 61970, and 61968), provide foundational aspects and principles but remain limited. As stated in IDSA [9] for energy interoperability documentation (towards ensuring compliance with the up-to-date guidelines as stated in the DoA), the H2020-SYNERGY project[10] advanced semantic harmonization through its Common Information Model, aligning prominent energy data models and defining their interrelationships. Future solutions require sector-specific Common Information Models as foundational harmonization tools, followed by orchestrated alignment between these models to enable unified smart energy systems.

Moreover, IDSA provides an overview of the technical standards and ontologies to ensure interoperability, security, and efficient data exchange among a wide variety of devices and platforms. Among the most critical standards are IEC 61850, which defines communication protocols and data models for intelligent electronic devices in substations; IEEE 2030.5 [15], which focuses on interoperability for energy management systems and smart meters; OpenADR [18], which standardizes demand response communications; and the IEC CIM family (61968/61970/62325)[11] and IEC 61850[12], which provides a comprehensive information model for managing utility operations and market transactions.

Beyond these, the ETSI SAREF[19] ontology, along with its extensions like SAREF4ENER and SAREF4GRID, provides standardized semantic models which enable better interoperability across energy and electricity grid domains. The international standardization process covers cloud computing and IoT and digital twin integration through ISO/IEC JTC 1/SC 38 and SC 41 which establish fundamental standards for data spaces and trusted data sharing and policy interoperability. Real-time data exchange in automation and building management systems depends on industrial communication protocols OPC UA[27] and Modbus which deliver secure and scalable frameworks with interoperable capabilities.

Together, these standards and ontologies form the backbone of smart grid and smart energy system integration, enabling interoperability, security, and efficient management of increasingly complex and heterogeneous digital infrastructures. In the context of the DIGITISE project, we first cluster the project needs to different categories towards the elicitation of standards for interest:

- Smartgrids and flexibility: while this is not the primary focus of DIGITISE project (as grid related aspects are not directly addressed in the project) the analysis is performed to ensure interconnection with the grid as well as flexibility services provision
- DERs (EVs/PVs/Batteries): as the decentralized energy assets that enable flexibility, resilience, and sustainability.
- Building environment: acting as active energy node in the DIGITISE project, integrating DERs and flexibility services.

The integration of smart grids and the enablement of flexibility services depend on several key data model standards that ensure interoperability, efficient information exchange, and coordinated control across the energy system. The CIM family constitutes the

foundational semantic model for smart grid interoperability. It covers power system management from generation through transmission and distribution, as well as market operations and asset management. CIM enables seamless data exchange between grid operators, market participants, and distributed energy resources (DERs), supporting both operational and market-based flexibility.

IEC 62746[16] standard represents a guideline which supports interface development between customer energy management systems and power management systems through its demand response and distributed energy resource integration capabilities. The IEC 62746-10- 1 follows the OPENADR structure to establish basic data protocols and service operations that connect electricity providers with their aggregators and end users through two-way information sharing. The standard enables automated demand response operations at customer sites by using open standards such as XML and IP for communication to execute load shifting and generation and storage management. The international standard IEC 62746-4:2024 establishes the systems interface between customer energy management system and power management system through its Demand Side Resource Interface specification. The IEC Technical Committee 57 published this standard to establish protocols which enable demand-side flexibility and resource participation by connecting customer energy management systems (CEMS) to grid and market operator systems for residential, commercial, and industrial sites. IEEE 2030.5 is also a widely adopted standard for the interoperability of energy management systems, smart meters, and DERs. It supports advanced energy management, demand response, and flexible grid operations by providing protocols for secure and standardized information exchange between grid operators and customer devices.

In the field of SAREF, SAREF4ENER and SAREF4GRID are the ETSI ontologies to provide standardized semantic models for energy and grid domains, enabling interoperability and integration of diverse energy-related information systems and devices. They are particularly relevant for the semantic alignment of flexibility services and device-level integration.

Moving beyond the smartgrids core, there are numerous standards to promptly model DER related aspects. IEC 61850-7-420 and IEEE 2030.5 are two of the most significant data model standards for distributed energy resources (DERs) covering different assets such as solar photovoltaic systems, battery storage, electric vehicles, and controllable loads. IEC 61850-7-420[13] defines an information model for (DERs) and distribution automation systems which belong to the IEC 61850. The standard outlines logical nodes together with data objects which describe the operational conditions and control functions and capabilities of DERs at both individual asset level and aggregated resource management through facility and microgrid energy management systems. The hierarchical and modular modeling approach of IEC 61850-7-420 allows utilities and operators to monitor and control and integrate DERs throughout distribution networks and even

transmit them into transmission networks. The standard allows various operational functions including voltage and frequency disturbance response and voltage and reactive power control and return-to-service logic while maintaining semantic interoperability between different DER technologies and vendors. The standardization process includes specific instances which focus on particular DER types.

IEC TR 61850-90-7 focuses on power converters in DERs in particular. In order to stabilize grid voltage and frequency, it specifies control interfaces for bidirectional power flow and Volt-VAR-Watt management functions. This enables DERs to dynamically modify reactive power and active power output. In order to coordinate thousands of assets and deliver grid-scale flexibility services, it also creates hierarchical control architectures for combining DERs into virtual power plants. A connection is made between IEC 61850 and electric vehicle (EV) charging protocols with the release of the IEC TR 61850-90-8 [14] technical report. This framework offers consistent communication between EVs, charging stations, and grid operators by mapping EV-specific standards like ISO/IEC 15118 (communication) and IEC 61851 (charging safety) to IEC 61850 semantics. Interoperability between various hardware, energy management systems, and utilities is guaranteed by this semantic alignment. Additionally, the open standard for communication between EV charging stations and central management systems is the Open Charge Point Protocol (OCPP [18]). An OCPP message is considered the "shared language" of the EV charging ecosystem. The "shared language" enables real-time messages about the status of the charging point, session start and end messages, billing metering data, firmware updates, diagnostics, and smart charging commands to be sent from or received by the charging station and/or central management system. The protocol has gone through a number of major version evolutions. OCPP 1.6, uploaded in 2015, incorporated smart charging, demand response, and load management capabilities and is still commonly seen in commercial installations. OCPP 2.0.1, uploaded in 2020, included many enhancements such as support for multiple connectors, improved security, improved transaction and device management, and improved user experience. OCPP 2.0.1 is the latest version published, with OCPP 2.1 to be published in 2025 including features for distributed energy resource (DER) control and vehicle-to-grid (V2G) capabilities and will be backwards compatible with OCPP 2.0.1. OCPP's international standard for communication between EV charging stations and management systems was cemented further in late 2024 when OCPP 2.0.1 Edition 3 was approved as IEC standard 63584. OCPP is the basis on which scalable, interoperable, and future-ready EV charging infrastructure will be created globally.

IEEE 2030.5 called Smart Energy Profile 2.0 aims to integrate distributed energy resources (DERs) with a communications protocol and application program interface. It follows a RESTful architecture and uses HTTP with XML (or EXI if used with smaller faster devices) for message encoding, allowing easy integration with modern IT and IoT systems. IEEE 2030.5 describes the roles of utility servers, aggregators, DER clients, and end devices.

Through the IEEE 2030.5 protocol utility, both parties can engage in secure two-way communications with each other and different types of DERs including: solar inverters, batteries, electric vehicle supply equipment and others. Under IEEE 2030.5, it provides remote monitoring, remote control and scheduling of DER operations as well as real-time control and event-driven control of DER operations for things like demand response signals or dynamic export/export limits.

Moving to the building environment, there are numerous efforts considering also the domain of interest and the relevant use cases. At BIM related aspects representation, Industry Standards such as ISO 19650[23], IFC (Industry Foundation Classes, ISO 16739[24]), and openBIM lead to the development of consistent, transferable digital models for standard use across different sectors and software platforms. These processes allow IoT, automation, and energy management systems to connect from the earliest design level of a building to efficient facility management, maintenance, and tracking of the facility lifecycle. Other approaches are the Brick Schema that represents a standardized, open-source ontology designed to unify the semantic description of physical, logical, and virtual assets within buildings. Developed as an RDF-based framework, Brick enables machine-readable representations of building subsystems—including HVAC, lighting, security, and sensors—along with their relationships. Unlike proprietary or domain-specific models, Project Haystack[21]: A tag-based schema for labeling time-series data from IoT devices (e.g., sensors, meters). While lightweight and easy to deploy, Haystack lacks Brick's formal [20]semantics for relationships, limiting complex reasoning. Building Topology Ontology (BOT)[22]: Defines spatial hierarchies (e.g., sites, buildings, floors, rooms) without detailing equipment. Brick integrates BOT for location context but extends it with asset-specific semantics.
Considering the progress on SAREF, ETSI SAREF4BLDG defines a common vocabulary and structured relationships for building automation systems, ensuring that data from heterogeneous devices (e.g., thermostats, luminaires, CO_2 sensors) can be interpreted uniformly across applications. It bridges gaps between proprietary systems, IoT platforms, and energy management tools. While the aforementioned schemas cover the top-down approach at building environment (focus on building environment representation), IoT related data models are also considered to model the information as gathered from the different building systems. These models are coming from the IoT initiatives and the most prominent are presented ISO/IEC 18012-1 establishes a framework for seamless interoperability between IoT sensors, actuators, and controllers in building environments. It defines standardized interfaces, data formats, and communication protocols to ensure plug-and-play integration across heterogeneous devices. MODBUS/ BACnet/KNX are globally adopted open protocol for building automation, supporting interoperability across HVAC, lighting, fire safety, and access control system. Matter is a smart home and IoT standard by the Connectivity Standards Alliance (CSA)[25] to enable seamless

interoperability and compatibility between smart devices from different manufacturers, eliminating the need for multiple proprietary hubs and ensuring that devices can work together reliably, securely, and locally within any Matter-certified ecosystem[25].

2.1.2 Health/ comfort/security domain standardization overview

In this section, the interest is to address the non-energy related aspects to be examined in the project. Starting with the comfort related aspects within the building environment, special reference to the work provided by IEA EBC - Annex 79 - Occupant-Centric Building Design and Operation [28][29]. The field of occupant modelling emerged over four decades ago; however, it has surged in the past decade – particularly as a result of IEA EBC Annex 66 – "Simulation and Definition of Occupant Behaviour in Buildings". Annex 66 played an important role in formalizing experimental research methods, modeling and model validation, and occupant simulation. The key principles of the relevant work are presented below and adapts the well-known DNAs principles.

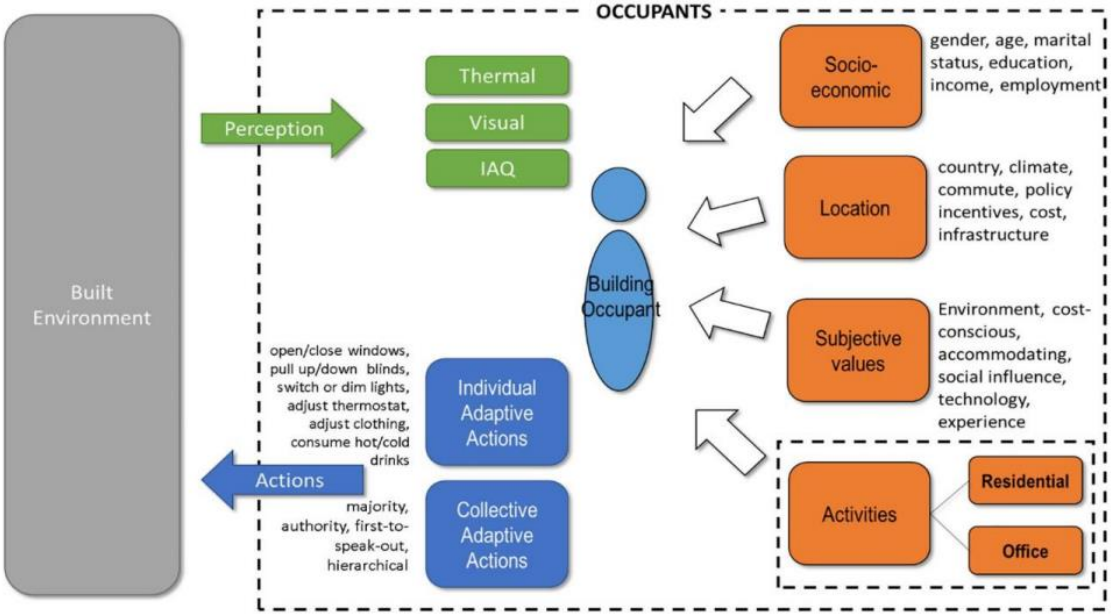

FIGURE 1 IEA EBC – ANNEX 79 – OCCUPANT-CENTRIC BUILDING DESIGN AND OPERATION OVERVIEW

In terms of modeling, the starting point of the work is the obXML schema[30] that stands as the basis for further work in ontological modeling. The obXML schema (occupant behavior XML) is a standardized XML schema, that can be deployed for representing and modeling occupant behaviors in buildings. The obXML schema is focused on representation and modeling of behaviors that impact building energy consumption. The obXML schema is driven by the DNAS (Drivers, Needs, Actions, Systems) heuristic that provides a more systematic way of describing the interconnectedness of occupants and building systems. Drivers are the motivators that compel occupants to perform behaviors (e.g. temperature, daylight, schedules). Needs are physical or psychological needs that need to be satisfied (e.g. comfort, privacy). Actions are interactions or activities that

occupants perform (e.g. opening windows, changing thermostat). Systems are components or assets in the building that occupants interact, e.g. HVAC, lighting, windows. A reference model interpretation of the obXML is the Digital Construction Ontologies (DiCon [32]) that form a modular, standards-aligned semantic framework designed to integrate data and knowledge across the full lifecycle of buildings—from design and construction to operation and renovation. Developed to bridge fragmented digital tools, DiCon enables interoperability, reasoning, and analytics in modern, data-driven construction environments. At its core, DiCon recognizes that digital construction is not just about static BIM models or isolated IoT data streams. It is also understanding the relationships between physical assets, human agents, activities, environmental variables, and contextual shifts over time thus addressing occupants' behaviour in the building environment following the DNAs principle. A similar approach from the Occupancy Profile ontological [1]model to handle the user's interaction within the building environment as well as the Occupant Feedback Ontology [31] to represent concepts related to human interaction within the building environment.
Apart from comfort related aspect, IAQ related aspects are covered from the concepts defined by this modeling approach as stated above ([32]) (while there are also IAQ specific attempts that aim to model data representation for Indoor Air Quality, Building Energy Performance, and Health Data Computation [33]). Nevertheless, the review of the relevant IAQ standardization is needed in order to define the relevant metrics/ boundaries of interest for the applications to be delivered in the project.

Standardized data models for indoor air quality (IAQ) (ISO 16000 Series has required IAQ management systems and measurement methods, and reporting for all indoor environments except industrial/agricultural environments, ASHRAE Standards 62.1 & 62.2[35] provide guidance on ventilation and IAQ in commercial and residential buildings, a common framework for use in North America and around the world, EN 15251, EN 16798 European standards that provide ventilation performance, simulation of its ventilation, etc.) are essential to consistent data capturing, sharing, and assessment across diverse platforms, organizations and regulatory authorities. These models enable interoperability, support regulatory compliance, and enhance the potential of advanced analytics supporting overall building health and occupant well-being, advancing toward measurable targets for healthy environments. ASHRAE 62.1 defines CO_2 concentrations to be limited to 1,000 ppm in occupied spaces, while the European standard EN 16798-1 allows (1,500 ppm) for lower-category buildings. Particulate matter (PM) must be <25 μg/m³ for PM2.5 and <50 μg/m³ for PM10 in a 24-hour average. Volatile Organic Compounds (VOCs) have limits for specific compounds while total VOC (TVOC) concentration concentrations should not exceed 500 μg/m³. Meanwhile, occupant-Centric KPIs are measures of

[1] https://bimerr.iot.linkeddata.es/def/occupancy-profile/

perceived air quality. ISO 16000-40 [37] includes occupant satisfaction surveys in which ≥80% of occupant acceptance defines "acceptable" air quality. Odor intensity scales (0-5) show the level of sensory pollution, where >2 odor intensity is viewed as problematic. Thermal comfort ranges (20-26°C operative temperature) in EN 15251 may also have influence on perceived indoor air quality. By defining the key model principles for health-related aspects within the building environment, will incorporate these domain specific attributes as part of the data model.

At the last part, building security related aspects may be incorporated as part of the model. Building security in the context of DIGITISE encompasses the protection of occupants within a facility (and not cybersecurity related aspects). The Industry Foundation Classes (IFC) schema is the prominent, international standard for Building Information Modeling (BIM), supporting a wide range of building elements and their properties. While IFC natively supports some aspects of security (such as SecurityRating-Index-based rating system indicating security level-, fire and acoustic ratings), its direct support for detailed physical security features—like burglary resistance, access control, and intrusion detection—is limited. However, IFC can be extended and adapted to address these needs through property sets (Psets), custom attributes, and ontology mappings. Namely, classes such as ThreatLevel, DelayTime, AttackMethod, AlarmIntegration may be considered in order to enhance the holistic data model with security related parameters.

Conversely, COBie is a widely adopted data schema for asset management and handover in construction projects. Although COBie does not contain explicit security rating fields, it does enable the use of custom attributes and performance data per the Employer's Information Requirements (EIR). Therefore, security rating information (such as resistance ratings and certification to security standards) could be part of the COBie supported handover, in order to meet facilities management and regulatory requirements.

2.1.3 Financial domain standardization overview

The focus in this section is to address aspects of interest in the DIGITISE project, namely related with projects financing as well as investments in the energy sector. Following the review of the relevant work the most promising model (domain agnostic) is the FIBO ontology. FIBO [38] is a conceptual ontology developed by the Enterprise Data Management Council (EDMC) to standardize the language of the financial industry. It provides precise definitions for financial instruments, business entities, and financial processes. Key FIBO modules include:

- Foundations (FND): This includes fundamental concepts that underpin the entire ontology, such as identifiers, arrangements, and general abstract concepts used across finance.
- Business Entities (BE): This covers the types of entities that exist within the financial system, including legal entities (corporations, partnerships, government entities), individuals, and the concepts of ownership and control.

- Finance Business & Commerce (FBC): This domain deals with general financial business practices and concepts, including:
    - Financial Instruments: Defining various types of financial instruments like agreements, contracts, notes, equities, options, and debt instruments.
    - Products: Describing financial products offered.
    - Loans & Mortgages: Specific concepts related to lending.
- Securities (SEC): This domain focuses on securities and related concepts, including:
    - Securities: Defining different types of securities.
    - Derivatives: Covering various derivative instruments.
    - Funds: Concepts related to investment funds.
    - Indicators & Indices: Defining financial indicators and market indices.
- Regulatory & Compliance (REG): Modules related to financial regulations, regulatory agencies, and compliance requirements.

These modules provide a common vocabulary to describe financial information, in order to help interoperability and data governance capabilities in the financial sector.

It should be noted that FIBO is at root a generic model as a comprehensive standard for financial concepts, and thus adaptation for project needs and project objectives seems necessary. A high-level elicitation process of concepts of interest and interconnections with the energy concepts, as presented above, is conducted:

- fibo-fbc-fi-fi:FinancialInstrument: The overarching concept for any financial asset. PV and battery projects, or their underlying financing structures, would be types of financial instruments.
- fibo-fbc-fi-fi:Loan: If the DER investment involves debt financing (e.g., bank loans for project development), concepts related to loans, interest rates, repayment schedules, and collateral would be highly relevant.
- fibo-fbc-fi-fi:Equity: If the investment involves equity participation (e.g., ownership shares in the project company), concepts related to equity instruments, share classes, and ownership structures would apply.
- fibo-fbc-fi-fi:Security: For more complex investment structures that might involve tradable securities (e.g., green bonds, project bonds), FIBO's extensive securities modules would be invaluable.
- fibo-fbc-fi-fi:Contract: The underlying agreements for power purchase (PPAs), energy storage services, or maintenance agreements would be instances of financial contracts.
- fibo-fbc-fi-fi:InterestRate: For calculating the cost of capital, discounting cash flows, and evaluating different financing options.

Considering the different business entities and roles:

- fibo-fbc-be-be:LegalEntity: Representing the various entities involved in the DER project: project developer, investor, Financial institutions, Manufacturers of PV panels or batteries, ESCO etc.
- fibo-fbc-rel-rel:PartyInRole: To define the specific roles these entities play within the investment structure (e.g., "Lender," "Borrower," "Investor," "Project Owner," "Energy Producer," "Energy Consumer").

Also considering Market Data & Pricing schemas:

- fibo-fbc-idm-idm:MarketData: While FIBO doesn't directly model electricity prices (covered by IEC standards mentioned above) or battery degradation rates, it provides the framework for what constitutes market data. This would be crucial for integrating external data feeds on:
    - Electricity Spot Prices: For revenue projections.
    - Forward Electricity Prices: For long-term PPA valuation.
    - Carbon Credit Prices: If carbon offsets are part of the revenue stream.
    - DER Equipment Costs: Prices for PV modules, inverters, batteries, etc.
    - Interest Rate Curves: For financial modeling.
- fibo-fbc-ind-ind:Indicator: FIBO's concepts around financial indicators could be extended or inspire how specific performance metrics (e.g., LCOE, NPV, IRR) are defined and related to the investment.
- fibo-fbc-fi-fi:Valuation: Although not providing specific algorithms, FIBO provides the conceptual basis for what a valuation is and what it applies to (e.g., the valuation of a project, a security, a contract).
- fibo-fbc-fi-fi:CashFlow: The concept of expected cash flows, which are fundamental to any investment sizing application.

It is evident that the FIBO model may stand as the basis for the work in DIGITISE projects. Aspects related to project financing, performance and purchase related parameters are covered by the relevant model.

An alternative to the FIBO model is the XBRL, which stands for eXtensible Business Reporting Language [39], as a global, open, and freely available standard for the electronic communication of business and financial data. At its core, XBRL is about giving structure and context to business information. Instead of treating financial reports as static blocks of text (like in a PDF or a printed document), XBRL provides unique, machine-readable "tags" for each individual item of data. For example, a taxonomy for financial reporting might define "Net Profit," "Total Assets," or "Revenue." Taxonomies ensure that data is consistently categorized and understood across different systems. They include various "linkbases" that define:

1. **Label Linkbase:** Provides human-readable labels for concepts (in multiple languages).
2. **Reference Linkbase:** Links concepts to authoritative accounting standards or regulatory rules.

3. **Calculation Linkbase:** Specifies arithmetic relationships between items (e.g., Assets = Liabilities + Equity), allowing for automated validation.
4. **Definition Linkbase:** Defines other relationships and constraints between concepts.
5. **Presentation Linkbase:** Dictates how elements should be presented in a report.

In relation to the project activities, XBRL is a rather generic approach though some linkage may be foreseen, by mapping the financial and key performance metrics to an XBRL taxonomy. More specifically, these are contextual or descriptive elements. The ProjectName could be included as part of a narrative disclosure or serve as an entity identifier. The InvestmentPeriod_years might function as an axis (dimension) for reporting period-specific facts within an XBRL instance, or it could be a numerical fact itself if explicitly disclosed. Standard taxonomies contain concepts for NetPresentValue, InternalRateOfReturn, and ReturnOnInvestment. Similarly, concepts like PropertyPlantAndEquipment_Cost (for CAPEX) and OperatingExpenses (for OPEX) are defined. General concepts such as Revenue, OtherIncome, and GovernmentGrants are also available. Furthermore, concepts about Debt, Equity, InterestRate, and CapitalContributions are established within existing financial taxonomies.
In essence, XBRL provides the *reporting envelope* and standardization for sharing the financial and key performance outcomes of a DER investment sizing application, making its outputs machine-readable and comparable. An instantiation of the different concepts and attributes to the project needs is required.
Apart from the aforementioned widely adopted models and standards, there are domain specific models to be considered. Special reference to specific market available solutions that have defined proprietary models to address the relevant aspects. The two most well-known applications are the HOMER [40]and REOpt[41] applications that are briefly presented (with focus on the data model aspects to be examined as relevant in DIGITISE project). HOMER Energy offers a suite of software tools—HOMER Pro, HOMER Grid, and HOMER Front—designed to model, optimize, and evaluate distributed energy systems and microgrids, from off-grid villages to utility-scale renewable projects. While the exact database schemas of commercial software are proprietary, their extensive documentation, input requirements, and the fundamental principles of energy project finance allow us to construct a detailed conceptual data model. This model serves as a blueprint for any application focused on Distributed Energy Resource (DER) asset sizing and investment, revealing how technical and financial data must be structured to produce reliable techno-economic analysis.
At its core, the data model for a tool like HOMER is not a single, flat table but a relational structure. The data model can be broken down into five primary entities: Projects, Components, Resources and Financials.

- The Project is the central entity that brings all other elements together for a specific analysis. It defines the scope and constraints of the simulation with attributes such as: ProjectID, ProjectName, ProjectType, ProjectLifetime
- The Component Entity: This is the most complex entity, representing the physical assets available for the system. It's often structured as a catalog of component types with specific instances being part of a project: ComponentID, ComponentType, CapitalCost, ReplacementCost, O&MCost etc
- The Resource Entity: This entity holds time-series data representing the available renewable energy at the project's location: ResourceID, Location, TimeSeriesData etc
- The Financials Entity: This entity defines the macroeconomic assumptions and financial structures that influence the project's investment viability: DiscountRate, InflationRate, TaxRate etc

Beyond the foundational model exemplified by HOMER, several other powerful applications offer distinct approaches to DER asset sizing and investment. Their data models, while sharing a core structure, are uniquely tailored to their specific strengths—be it detailed financial analysis, electrical system fidelity, or resilience planning. The NREL's REopt developed by the National Renewable Energy Laboratory (NREL), is a techno-economic optimization platform focused on helping users identify the optimal mix of DERs to achieve cost savings, clean energy goals, and enhance energy resilience. Its data model is heavily weighted towards financial performance and outage simulation.

- Emphasis on Financials and Utility Tariffs: REopt's model is a sophisticated in its representation of electricity costs. It includes a deep, structured entity for UtilityTariff that goes far beyond a simple energy rate.
    - UtilityTariff Entity: Contains attributes for energy charges (EnergyRates), demand charges (DemandRates), fixed charges, and tiered rate structures.
    - DemandRates Attributes: Further broken down by Time Of Use (TOU) periods (e.g., on-peak, mid-peak, off-peak), seasons, and whether they are based on coincident or non-coincident peaks.
    - Incentive Granularity: The Financials entity has robust support for a wide array of federal and state incentives, each with its own set of rules and caps.
- The Resilience Entity: REopt's data model explicitly includes entities and attributes to support resilience analysis.
    - OutageSimulation Entity: Defines parameters for a hypothetical grid outage, including OutageDuration (in hours) and the CriticalLoad profile that must be sustained.
    - CriticalLoad Entity: A separate time-series load profile, distinct from the main site load, representing mission-critical energy needs. The optimization goal can be set to "maximize resilience" by ensuring this load is met.

- Simplified Component Model: Compared to more engineering-focused tools, REopt's Component model is slightly more abstracted. It focuses on the economic and high-level performance attributes (CAPEX, OPEX, efficiency, lifetime) rather than deep electrical characteristics.

It is evident that the FIBO model can stand as the base for the modeling work in DIGITISE project. Additional attributes from XBRL as well as the domain specific models of the commercial solutions may be considered to complement features not covered by the FIBO model adaptation in the project.

2.2 Review of data landscape

In this section, the review of the data landscape available at the demo sites as well as the data needs of the project applications is performed. The available data assets are derived from the preliminary work in T6.1 and the asset landscape analysis while the data needs are derived from the work in T2.4 and the elicitation of data assets in D2.2.
Starting with the review of the available physical assets (and the derivative data assets), of the project demonstrators and their respective assets in Greece, Spain, Croatia and Ireland. The focus is on the following areas:

- AMI & Digital Tools: Information on smart meter data, Advanced Metering Infrastructure (AMI) functionality, and other digital tools that gather aggregate energy data at building/smart meter level. Special reference to the IR demo site where the actual metering data are to be made available in the project from the provider
- Local Energy Communities (LEC): Community energy resources, covering Distributed Energy Resources including Photovoltaics (PVs), Electric Vehicle (EV) Charging Points (CPs), and batteries.
- Building & BEMS Assets: Building Energy Management Systems (BEMS) integration, Internet of Things (IoT) interaction, and relevant measurements.

This section details the specific data required to build and operate the various energy services within the DIGITISE project. These requirements, summarized in the table below, are derived from the preliminary design specifications for the DIGITISE energy applications as documented in D2.2. The scope is further expanded to include additional data requirements identified by the analytics services within the DIGITISE framework.

Table 2: Key Data Categories at the DIGITISE Applications

Energy Application	Key Data Categories
Energy Management and Self-consumption Optimization	Data from smart meters, electricity tariffs, customer and building profiles, renewable energy data, weather data, storage data, EV charge data, smart devices in buildings, weather, generation, demand and flexibility predictions

Energy Application	Key Data Categories
Smart Home and DER Automated Control Application	Data from IoT/smart meters, environmental monitoring within buildings, smart devices in buildings, DR events
Flexibility Virtual Power Plant Configuration Application	building IoT-specific smart meter data, battery system data, EV charging station data, flexibility profiles, asset configuration data
Open Flexibility Trading Marketplace and Smart Flexibility Contracts	building IoT-specific smart meter data, battery system data, EV charging station data, flexibility profiles, asset configuration data
Health and Comfort Application	Data from IoT meters, customer and building profiles, weather data, environmental monitoring within buildings, smart devices in buildings, comfort profiling, weather
Energy Behavioural profiling application	Data from IoT/smart meters, electricity tariffs, customer and building profiles, weather data, environmental monitoring within buildings, smart devices in buildings, comfort profiling, weather
RES and DER Investment Sizing and Guidance	Community evolution, smart meter data, data from renewable energy sources, data from EV charging points, data from storage systems, weather data, generation and demand forecasts
Personalized Prosumer Engagement and Capacity Building	Data from IoT/smart meters, electricity tariffs, customer and building profiles, renewable energy data, weather data, environmental monitoring within buildings, smart devices in buildings, comfort profiling, weather, generation and demand predictions

The development of the DIGITISE common data model is guided by the specific data assets and needs of our demonstration partners. We have compared these requirements with existing standards and models to verify their alignment with the project's goals. Through a subsequent data harmonization task, we will incorporate essential concepts from these standards into our common data model. This ensures complete coverage of the data requirements and availability expressed by the demo partners and application developers at this stage.

2.3 DIGITISE data model overview

Following the review analysis of the data models as well as the review of the data landscape, the elicitation of the key concepts and attributes is provided to support the

definition of DIGITISE data model. A high-level overview is provided in this section, considering also the detailed presentation of the model in Annex I.

As a starting point is the segmentation of the different concepts to the 3 core domains presented above: energy, health and finance considering the key focus on the 1st domain as of interest in DIGITISE.

Starting with the energy domain, the different sub categories are defined:

- Generation/ Storage Systems: This domain covers generation/ storage at the community/portfolio level, including renewable sources and monitoring systems for the different DER systems
- EV Charging & e-Mobility: This domain focuses on the entire electric vehicle charging ecosystem, from the physical infrastructure to the management systems and user roles.
- Buildings & IoT: This domain encompasses the physical structure of buildings, their internal climate and electrical systems, and the IoT devices that enable smart control and monitoring.
- Cross-Cutting Roles & Systems: This domain includes concepts that are fundamental across multiple domains, such as energy management systems and user roles that are not specific to a single area.

An overview of the key concepts defined as part of the models is provided below:

ID	Concept	Category
1	AirConditioningSystem	Buildings & IoT
2	AirConditioningSystemControlAction	Buildings & IoT
3	Boiler	Buildings & IoT
4	BoilerControlAction	Buildings & IoT
5	Building	Buildings & IoT
6	BuildingMeasurements	Buildings & IoT
7	BuildingSpace	Buildings & IoT
8	BuildingStorey	Buildings & IoT
9	BuildingZone	Buildings & IoT
10	ChillerDevice	Buildings & IoT
11	ChillerDeviceControlAction	Buildings & IoT
12	DomesticHotWaterSystem	Buildings & IoT
13	DomesticHotWaterSystemControlAction	Buildings & IoT
14	ElectricAppliance	Buildings & IoT
15	ElectricApplianceControlAction	Buildings & IoT
16	Gateway	Buildings & IoT
17	HeatPump	Buildings & IoT

18	HeatPumpControlAction	Buildings & IoT
19	LightingDevice	Buildings & IoT
20	LightingDeviceControlAction	Buildings & IoT
21	Outlet	Buildings & IoT
22	SmartAppliance	Buildings & IoT
23	SmartApplianceControlAction	Buildings & IoT
24	SpaceHeatingDevice	Buildings & IoT
25	SpaceHeatingDeviceControlAction	Buildings & IoT
26	Address	Cross-Cutting Roles & Systems
27	Aggregator	Cross-Cutting Roles & Systems
28	AggregatorPortfolio	Cross-Cutting Roles & Systems
29	BalancingResponsibleParty	Cross-Cutting Roles & Systems
30	BalancingServiceProvider	Cross-Cutting Roles & Systems
31	DemandResponseEvent	Cross-Cutting Roles & Systems
32	DemandResponseEventSignal	Cross-Cutting Roles & Systems
33	DemandResponseReport	Cross-Cutting Roles & Systems
34	DemandResponseReportReading	Cross-Cutting Roles & Systems
35	Device	Cross-Cutting Roles & Systems
36	DeviceControlEvent	Cross-Cutting Roles & Systems
37	DeviceControlEventAction	Cross-Cutting Roles & Systems
38	DeviceControlStatus	Cross-Cutting Roles & Systems
39	EnergyDemandMeasurements	Cross-Cutting Roles & Systems
40	EnergyMarket	Cross-Cutting Roles & Systems
41	EnergyMarketOperator	Cross-Cutting Roles & Systems
42	EnergyServiceCompany	Cross-Cutting Roles & Systems
43	Event	Cross-Cutting Roles & Systems
44	FacilityManager	Cross-Cutting Roles & Systems
45	Flexibility	Cross-Cutting Roles & Systems
46	FlexibilityMarket	Cross-Cutting Roles & Systems
47	FlexibilityMarketOperator	Cross-Cutting Roles & Systems
48	Incident	Cross-Cutting Roles & Systems
49	IncidentLog	Cross-Cutting Roles & Systems
50	KeyPerformanceIndicator	Cross-Cutting Roles & Systems
51	KeyPerformanceIndicatorValue	Cross-Cutting Roles & Systems
52	LoadResponse	Cross-Cutting Roles & Systems
53	LocalEnergyCommunity	Cross-Cutting Roles & Systems
54	LocalEnergyCommunityPortfolio	Cross-Cutting Roles & Systems
55	Location	Cross-Cutting Roles & Systems
56	Measurement	Cross-Cutting Roles & Systems
57	MeteringSystem	Cross-Cutting Roles & Systems
58	NetworkSwitch	Cross-Cutting Roles & Systems
59	Offer	Cross-Cutting Roles & Systems
60	OfferOption	Cross-Cutting Roles & Systems
61	Order	Cross-Cutting Roles & Systems

62	Period	Cross-Cutting Roles & Systems
63	Prosumer	Cross-Cutting Roles & Systems
64	Request	Cross-Cutting Roles & Systems
65	Retailer	Cross-Cutting Roles & Systems
66	RetailerPortfolio	Cross-Cutting Roles & Systems
67	Schedule	Cross-Cutting Roles & Systems
68	Settlement	Cross-Cutting Roles & Systems
69	Status	Cross-Cutting Roles & Systems
70	TroubleTicket	Cross-Cutting Roles & Systems
71	VirtualPowerPlant	Cross-Cutting Roles & Systems
72	WeatherMeasurement	Cross-Cutting Roles & Systems
73	WeatherStation	Cross-Cutting Roles & Systems
74	chargePointOperator	EV Charging & e-Mobility
75	chargePointOwner	EV Charging & e-Mobility
76	chargeSession	EV Charging & e-Mobility
77	ElectricVehicle	EV Charging & e-Mobility
78	EVChargingPlatform	EV Charging & e-Mobility
79	EVChargingStation	EV Charging & e-Mobility
80	EVChargingStationControlAction	EV Charging & e-Mobility
81	EVUser	EV Charging & e-Mobility
82	Battery	Generation/Storage Systems
83	BatteryControlAction	Generation/Storage Systems
84	BiomassPlant	Generation/Storage Systems
85	CombinedHeatingPowerSystem	Generation/Storage Systems
86	Electrolyzer	Generation/Storage Systems
87	EnergyGenerationMeasurements	Generation/Storage Systems
88	EnergyStorageMeasurements	Generation/Storage Systems
89	FuelCell	Generation/Storage Systems
90	Generator	Generation/Storage Systems
91	HydrogenTank	Generation/Storage Systems
92	HydroPowerSystem	Generation/Storage Systems
93	Inverter	Generation/Storage Systems
94	PhotovoltaicSystem	Generation/Storage Systems
95	PlantOperator	Generation/Storage Systems
96	PowerPlant	Generation/Storage Systems
97	RenewableEnergySource	Generation/Storage Systems
98	RenewableEnergySourceOperator	Generation/Storage Systems
99	SolarThermal	Generation/Storage Systems
100	SolarThermalControlAction	Generation/Storage Systems
101	WindTurbine	Generation/Storage Systems

TABLE 1 ENERGY DATA MODEL OVERVIEW

Note: While DIGITISE project is not addressing aspects related to the grid, some key concepts required for grid modeling may be considered for persistency reasons. Namely

concepts related to grid modeling such as: ACLine, Branch, Bus, ConnectivityNode, DistributionSystemOperator, Impedance, Load, SCADA, Substation are defined.

Moving to the health/comfort/security sector, there are different

- Occupant Comfort & Behavior: This domain focuses on modeling the dynamic interaction between building occupants and their environment. It utilizes DNAS (Drivers, Needs, Actions, Systems) principle, to define the relevant concepts
- Indoor Air Quality (IAQ) & Health: This segment addresses the measurement, management, and standardization of indoor air quality to ensure occupant health and well-being.
- Physical Building Security: This domain focuses on integrating physical security concepts into the building data model, distinct from cybersecurity. The reason for this is the fact that the health aspects will be complemented by the security related aspects to be examined in the project.

An overview of the key concepts defined as part of the model is provided below:

ID	Concept	Category
1	Sensor	Cross-Cutting Roles & Systems
2	SensingMeasurement	Cross-Cutting Roles & Systems
3	HealthProblem	Indoor Air Quality (IAQ) & Health
4	Humidifier	Indoor Air Quality (IAQ) & Health
5	HumidifierControlAction	Indoor Air Quality (IAQ) & Health
6	IndoorAirQuality (IAQ)	Indoor Air Quality (IAQ) & Health
7	IndoorEnvironmentHealth	Indoor Air Quality (IAQ) & Health
8	MedicalDevice	Indoor Air Quality (IAQ) & Health
9	MedicalDeviceControlAction	Indoor Air Quality (IAQ) & Health
10	PollutantConcentration	Indoor Air quality (IAQ) & Health
11	VentilationSystem	Indoor Air quality (IAQ) & Health
12	VentilationSystemControlAction	Indoor Air quality (IAQ) & Health
13	AvoidableDisposition	Occupant Comfort & Behavior
14	Occupancy	Occupant Comfort & Behavior
15	OccupancyActivity	Occupant Comfort & Behavior
16	OccupancyProfile	Occupant Comfort & Behavior
17	Occupant	Occupant Comfort & Behavior
18	OccupantBehavior	Occupant Comfort & Behavior
19	SoundComfort	Occupant Comfort & Behavior
20	SoundLevel	Occupant Comfort & Behavior
21	ThermalComfort	Occupant Comfort & Behavior
22	VisualComfort	Occupant Comfort & Behavior
23	AccessControl	Physical Building Security
24	AlarmIntegration	Physical Building Security

25	AttackMethod	Physical Building Security
26	BurglaryResistance	Physical Building Security
27	IntrusionDetection	Physical Building Security
28	Physical Security	Physical Building Security
29	ThreatLevel	Physical Building Security

TABLE 2 COMFORT, HEALTH & SECURITY DATA MODEL OVERVIEW

Then, the different segments for the financial related aspects are considered:

- Business Roles: This category defines the various stakeholders and their specific functions within the business ecosystem. It clarifies the responsibilities, relationships, and interactions of key participants involved in energy and financial projects
- Financial Instruments: This category covers the specific financial assets and contractual instruments that were utilized to finance, own and operate projects. Examples include foundational concepts such as Loans (debt financing), Equity (ownership interests), and Securities (green bonds); along with a Contract like Power Purchase Agreements (PPAs) that provides for revenue and service delivery.
- Financial Indicators: This category includes the key performance indicators (KPIs, InterestRate, ProjectValuation) commonly used to assess and evaluate the financial viability and performance of an investment.
- Utility Tariffs: The utility pricing structure for electricity is included in this category, because it is essential for conducting economic analyses of energy projects.

An overview of the key concepts defined as part of the model is provided below:

ID	Concept	Category
1	Arrangement	Agreements & Contracts
2	Contract	Agreements & Contracts
3	Lease	Agreements & Contracts
4	Loan	Agreements & Contracts
5	Mortgage	Agreements & Contracts
6	RepaymentSchedule	Agreements & Contracts
7	FinancialIndicator	Financial Instruments & Valuation
8	FinancialInstrument	Financial Instruments & Valuation
9	FinancialProduct	Financial Instruments & Valuation
10	InterestRate	Financial Instruments & Valuation
11	ProjectValuation	Financial Instruments & Valuation
12	Security	Financial Instruments & Valuation
13	FinancialBody	Project Roles & Entities
14	Investor	Project Roles & Entities
15	Lender	Project Roles & Entities

16	Manufacturer	Project Roles & Entities
17	Project	Project Roles & Entities
18	ProjectDeveloper	Project Roles & Entities
19	FinancialRegulation	Regulatory
20	RegulatoryAgency	Regulatory
21	Incentive	Utility, Tariffs & Incentives
22	TariffProfile	Utility, Tariffs & Incentives
23	TariffRateComponent	Utility, Tariffs & Incentives
24	UtilityBill	Utility, Tariffs & Incentives
25	UtilityTariff	Utility, Tariffs & Incentives

TABLE 3 FINANCE DATA MODEL OVERVIEW

By presenting the details of the model, the key tech principles considered for the management of the model are presented below.

2.4 Semantic Interoperability Management

The task of Semantic Interoperability Management provides a means of dealing with heterogenous datasets with semantic alignment in the DIGITISE Data Model. This layer provides the tools and methods needed to confirm that the project data model retains its meaning and context, as it pertains to the creation, update, or deprecation of concepts within your model. This semantic maintenance allows us to provide interoperability in the most seamless manner, and enable the management of the DIGITISE data value chain, with the eventual reuse and trustworthy interpretation by all stakeholders. The key features of the Semantic Interoperability Management are as follows:

- Semantic Interoperability Management: The DIGITISE Data Model has an administrator (CIM Administrator) who can create, change, delete, concepts, fields, and their relationships. All users can access the model and submit suggestions and can also make suggestions for changes to the model, and the administrator can approve or deny the suggestions.
- Metadata and Version Control: The component utilizes a structured metadata schema to manage all model elements and ensure alignment with industry standards. A robust version control system tracks all major and minor modifications, guaranteeing consistency as the data model evolves.
- Semantic Model Repository: This component is a centralized system designed to store, manage, govern, and serve the DIGITISE semantic model. It acts as a "single source of truth" for the definitions, relationships, and vocabularies that describe a domain of interest in DIGITISE.

The Semantic Interoperability backend is a NodeJS application, with structured data storage through PostgreSQL, and enhanced by Elasticsearch for full-text search capabilities across the model. A synchronization process keeps the Elasticsearch indices synchronized with the main database. The modern referenced front-end provides role-

based access to model managers and data providers, and is developed with Vue.js and Tailwind CSS to maintain, reactive and intuitive user experience.

3 Data Collection and Governance

The Data Collection & Governance Layer takes care of the initial data intake process for DIGITISE according to D2.2. The layer provides an efficient system to gather data from different sources while maintaining proper curation and secure structured storage. This foundational layer establishes essential standards to maintain data quality and consistency together with governance compliance. The DIGITISE platform achieves data validation and traceability along with processing stage requirements through this layer. The following section provides detailed information about the fundamental components which establish the data collection and governance layer.

3.1 Data Collection Component

3.1.1 Overview

The Data Collection module serves as the central engine for executing data ingestion according to predefined configurations, ensuring that the specified processes are adhered to. It orchestrates its operations through various sub-components; each tailored to different data acquisition strategies. These strategies include (i) handling batch uploads for efficiently processing large data volumes at scheduled times, (ii) subscribing directly to real-time data streams using PubSub messaging for continuous, low-latency ingestion, and gathering information via both standard and (iii) custom APIs to accommodate structured and unstructured data from numerous external sources.
It also collects data from the DIGITISE IoT and DER connectors, which could be from Modbus, thus, supporting connectivity with industry-specific communications. In this manner, Data Collection uniquely offers a flexible, top-tier solution for collecting varied data functions, making it the foundation for sound data processing and analytics.

3.1.2 Delivered Functionality

In this section, we present the implemented functionalities of the component as implemented in the initial release in M15, aligned with the requirements identified in D2.1 and the detailed functional design IN D2.2. The following core features have been delivered:

- **Batch Data Collection (DC-01):** The system is capable of efficiently handling large volumes of data accumulated over time. Through configurable parameters, it determines the appropriate data sources and transformation rules, ensuring reliable and scalable batch ingestion.
- **On-Demand Data Collection (DC-02):** This functionality enables dynamic triggering of API calls to retrieve data from both third-party sources and APIs available within the DIGITISE data space. Scheduling of the API retrieval process is also supported in this 1st version

- **Streaming Data Collection (DC-03):** To supports streaming data ingestion. It leverages PubSub protocols such as MQTT to subscribe to real-time data streams, ensuring also that type of data integration.
- **Metadata Management (DC-04):** The component manages and updates metadata from the configuration and ingestion process. This ensures accurate tracking and documentation of data lineage and transformations from the outset.

All in all, the different features as defined in D2.2 have been partially implemented in the 1st version of the Data Space and made available for early testing and demonstration. Integration with DER connectors (such as MODBUS and OCPP) is not included in this initial release. These functionalities are planned for implementation in the upcoming second release. Also, the real time data streaming functionality provided in this version is a basic version implementation.
By delivering the above features (DC-01, DC-02, DC-03, DC-04), the system ensures controlled, efficient, and scalable data ingestion, forming a robust foundation for the overall data integration strategy and supporting reliable data flow throughout the ecosystem.

3.1.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- Within the cloud environment, specifically the DIGITISE Data Platform, different data ingestion methods are fully supported, allowing for maximum flexibility and integration. However, when operating with 3rd party API data ingestion, partial support to the authentication methods to be considered for connectivity.
- Regarding DER connectors, as stated above, these are not implemented in full in the current version and thus will be incorporated in the 2nd version.
- Regarding metadata configuration, a subset of the available metadata is supported. In the 2nd release, the configuration process will be enhanced in order to adhere the additions required in the project.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for users planning their data integration and monitoring strategies. In the next release, the aim is to enhance the existing features in order to ensure full support of the data collection process.

3.2 Data Harmonization Component

3.2.1 Overview

The Data Harmonization component serves as the core engine for transforming raw input data into the standardized DIGITISE Data Model as reported in Section 2. It executes

mapping operations according to predefined configurations, ensuring seamless conversion of source data into a structured format optimized for downstream processing. Upon receiving data from upstream sources, the component processes each record through systematic field translation, converting source formats into the DIGITISE Data Model's structure. This involves:

- Direct field-to-field mapping
- Data type conversions
- Contextual enrichment with metadata

The result of this task is harmonized data that may be consumed by the downstream applications and analytics workflows for the DIGITISE functions.

3.2.2 Delivered Functionality

In this chapter we describe the capabilities of the Data Harmonization component delivered in the first release, based on the requirements logged in D2.1and the full functional design in D2.2. In this release, the only functionality delivered relates to semantic mapping (SM-01). In this beta version, the Data Mapper has been fully implemented to semi-automatically convert raw and heterogeneous data into the standardized DIGITISE Data Model. The Data Mapper successfully applies semantic mapping to incoming data by independently applying a set of mapping rules and mapping logic, to determine how each record resulting from each processing operation aligns to the defined structure (target schema). In this first release, the Data Harmonization component allows for rigorous field to field mapping, conversion between data types, as well as the augmentation of records with additional, relevant metadata. Furthermore, full schema validation processes have been delivered, which ensures that all data processed is assured to comply with high levels of integrity standards and subsequent downstream processing. The gross application structure, implementation, stocking mechanisms, scheduled backup approaches and security, as articulated in the release represent a firmly established, scaled and secured architecture from which to continue to build future substantive and potential innovative extensions to data management, within the expected potential of the DIGITISE ecosystem.

3.2.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While data transformation is fully supported, the incorporation of additional transformation rules may be applied in order to ensure full coverage of the different transformations
- As already stated, the semantic model will be continuously improved through the life of the project to respond to emerging needs and requirements. Accordingly, any adjustment or change in the data model will also be addressed in the semantic harmonizer component. This will ensure that the data harmonization cycle can

remain fully aligned to the most recent specifications of the arrange of data model, which is necessary for consistency and reliability over use cases within a system. Each time there is an improvement to the semantic model, the harmonizer will detect schema reference changes as affectively and if it accepts them, it will apply the appropriate mappings, transformations, and reject or correct the check of transformed data under the revised model.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for users planning their data harmonization process. In the next release, the aim is to enhance the existing features in order to ensure full support of the data harmonization.

3.3 Data Curation Component

3.3.1 Overview

The Data Curation Engine is an integral piece that implements the quality improvement step across the data management lifecycle. Its goal is to apply the data curation processes specified during design during the project lifecycle. As it receives harmonized data, it carries out base data validation, cleansing and enrichment. In harmonizing the data, the Data Curation Engine will discover inconsistencies, missing values, duplicates and any other errors that will affect the usability and validity of the data. Similarly, the Data Curation Engine utilizes configurable rule sets to implement different usage cases around data quality and can adjust dynamically to multiple scenarios with different requirements. The Data Curation engine is therefore a flexible and scalable data curation and enrichment solution.

3.3.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Curation Engine as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. In this initial version, the primary delivered functionality is the handling of data curation. The Data Curation Engine has been implemented to apply a range of data quality processes to the standardized data. It reliably improves data quality by applying predefined validation and cleansing logic, ensuring each record meets the required quality standards.

The first release supports robust data validation against defined standards rules, cleansing of records to handle outliers and duplicates, and data enrichment where applicable. Comprehensive rule-based validation has also been delivered, ensuring that all processed data maintains high integrity and is fit for purpose.

3.3.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While core data cleansing and validation processes are fully supported, the incorporation of more advanced or domain-specific quality rules may be applied in the future to cover specialized use cases and enhance the overall quality assessment.
- As previously mentioned, the data quality standards and rule sets will be continuously updated throughout the project to address emerging needs. In line with this approach, any changes made to the quality rules will also be reflected in the Data Curation Engine.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for users planning their data curation process. In the next release, the aim is to enhance the existing features in order to ensure full support of data curation.

3.4 Data Storage Component

3.4.1 Overview

As one of the fundamental layers of the DIGITISE architecture, the Data Storage layer allows for data management for everything ingested, in a secure and high-performing way. It utilizes a multi-database strategy that considers relational, NoSQL, and graph technologies to accommodate structured, semi-structured, and unstructured data. It is also designed for scalability, supporting a distributed architecture that will enable handling massive amounts of data while maintaining high availability and low latency; this way users can efficiently process and access data. This provides a strong data basis for all data-related activities in the ecosystem.

3.4.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Storage layer as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. In this initial version, the core storage entity has been implemented to support the entire data lifecycle within the project. The analysis is presented considering the different storage elements delivered in this release, covering:

- Ingested Data (DS-01): Secure storage for all structured and unstructured data collected from various sources within the data space.
- Metadata Storage (DS-02): A dedicated repository for contextual information describing the datasets, which enables efficient data discovery, governance, and usage.
- Cross-Sector Vocabularies (DS-03): Storage for standardized terminologies and ontologies that facilitate data harmonization and interoperability across different industries (data models as defined in Section 2 of this deliverable).

The layer's scalable and high-performance design, as realized in this release, establishes a solid foundation for future data growth and evolving analytical needs within the DIGITISE ecosystem.

3.4.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While the core multi-database architecture is operational in the 1st version, further performance tuning and optimization for specific query patterns will be an ongoing activity as the data volume increases.
- As the project evolves, data lifecycle management approaches will be continuously refined to ensure both performance and cost-effectiveness. In line with this, the Data Storage layer will adapt to handle data at all stages.

In the next release, the aim is to enhance the existing features in order to ensure full support of data storage and management at scale.

3.5 Technology Stack and development details

Following the detailed presentation of the different components that consist of the Data Collection and Governance Layer of DIGITISE, an overview of the different technologies and frameworks considered for the delivery of the services is presented below.

Library	Version	License
NodeJS	18	MIT
FastAPI	0.110	MIT
Vue.js	2.7	MIT
TailwindCSS	2	MIT
Pandas	1.4.4	New BSD
NumPy	1.26	BSD
Kafka	-	Apache License 2.0
Zookeeper	-	Apache License 2.0
EMQX	-	Apache License 2.0
Apache Spark	3.3.4	Apache License 2.0
PostgreSQL	-	PostgreSQL License (similar to BSD/MIT)
Redis	-	3-clause BSD License
MinIO	-	Apache License 2.0
MongoDB	-	Apache License 2.0

TABLE 4 DIGITISE COLLECTION AND GOVERNANCE TECHNOLOGIES

In more details:

- The Data Collection component leverages a suite of technologies designed for robust and scalable data ingestion. For data streaming specifically, EMQX is employed as an MQTT broker to handle device communications. The service logic and connectors that manage these streams are built using backend technologies like NodeJS or the Python-based FastAPI.
- The Data Mapping service uses powerful data processing engines to turn raw data into a standard model. For large and distributed transformations, processing power is afforded by Apache Spark. For smaller, less complex mapping of data, Pandas (and other libraries) use Python to provide powerful data manipulation, usually depending on a backend service using FastAPI.
- The Data Curation component mainly utilizes a Python-based stack for data quality, cleansing, and validation. Formulating its analytical and manipulation capacity is NumPy for performing numerical logic, and Pandas for structuring and cleaning. These processes typically reside inside a backend service designed with FastAPI.
- The Data Storage layer employs a multi-modal strategy to efficiently handle diverse data types. PostgreSQL is used as the primary relational database for structured data, configurations, and metadata. For unstructured data and large data files (forming a data lake), MinIO provides S3-compatible object storage. MongoDB serves as the NoSQL document store for flexible, semi-structured data, and Redis is used as a high-speed caching layer to improve data access times.
- Any logging and communication among the different services is performed using Kafka brokers and Zookeeper as the coordination service.

Along with the presentation of the different open-source technologies considered for the delivery of the DIGITISE Collection and Governance layer, indicative screenshots from the 1st version of the application are provided to showcase the development progress on the work performed during the reporting period.

Note: we need to point out that basic screenshots are made available as part of the deliverable. The details will be made available through the platform documentation (available upon registration in the DIGITISE platform).

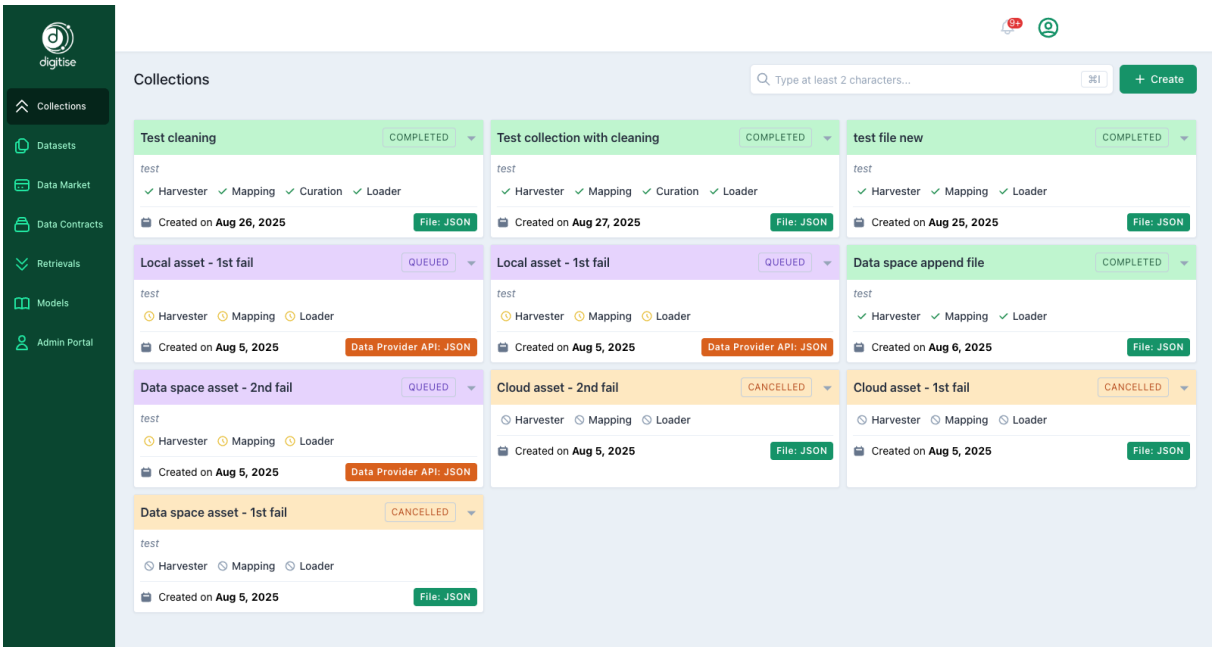

FIGURE 2 LIST OF DATA COLLECTIONS

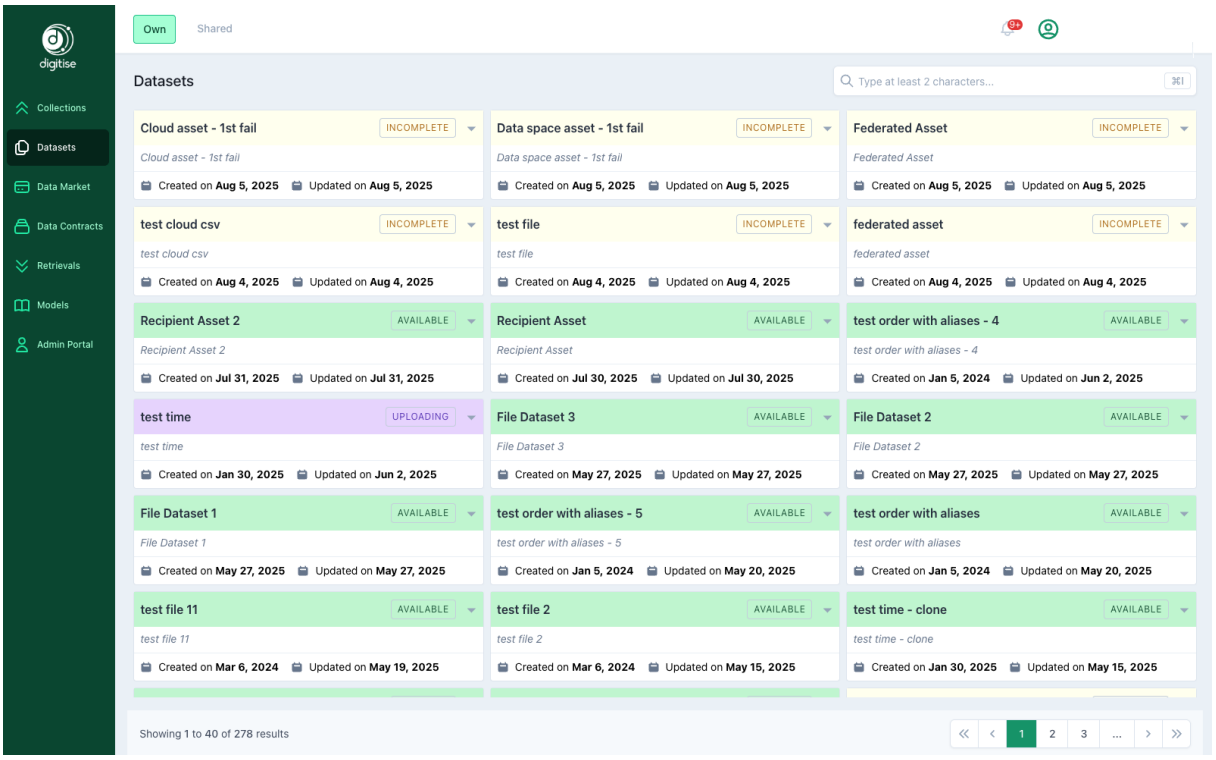

FIGURE 3 LIST OF DATA ASSETS

4 Data Privacy and Sovereignty

As outlined in the deliverable D2.2, the Data Privacy and Sovereignty Layer focuses on the security, management, and use of data within the DIGITISE community space, providing not only security and privacy to data but also increasing the sovereignty of data owners. For example, privacy can be created through proven usage policies, securing access, and transparency where participants have oversight and control of their aid data collection, processes, and sharing. As such, the purpose of the Data Privacy and Sovereignty Layer is to remove distrust within the ecosystem, through providing assurances that individuals retain sole authority over their contributions. The following provides background to the areas associated with the layers as described in D2.2.

4.1 Data Anonymizer Component

4.1.1 Overview

The Data Anonymizer Component is an important security feature of the DIGITISE ecosystem, and its purpose is to protect sensitive information in a way that complies with privacy law. The Data Anonymizer provides the data provider with a flexible way of stating and enforcing their anonymization rules. When the data anonymization component is implemented, the data provider can indicate the data fields needing protection, and then select from a wide variety of privacy-enhancing technologies, such as k-anonymity. After the appropriate anonymization techniques are configured, the Data Anonymizer will consistently apply the k-anonymization at runtime, so even if the data provider has indexed their data to protect and/or anonymize their raw data, in their anonymization configuration, before the data is locked down for analysis or shared with partners, in an acceptable manner, complying with privacy law.

4.1.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Anonymizer Component as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. In this initial version, the primary delivered functionality is the handling of data anonymization (DAC-01 as the sole feature for this component). The Data Anonymizer has been partially implemented to enforce specific strategies defined by data providers during the design phase. It reliably processes datasets by applying the configured anonymization techniques, ensuring sensitive fields are appropriately masked or generalized before the data is made available for downstream use.

The first release supports the application of rule-based anonymization, allowing providers to select specific fields and techniques to apply. The core logic has been implemented introducing basic anonymization techniques to be examined in the project.

4.1.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While basic anonymization techniques are supported, the initial release includes a specific set of methods. The incorporation of additional techniques may be explored in future releases to provide a wider range of privacy guarantees.
- The component is designed to be flexible, allowing data providers to refine their strategies to ensure the output data. In the current version there is no support on updating the rules initially applied during the configuration phase.

These constraints define the scope of functionalities available in the current version of DIGITISE while in the last release the aim is to enhance the existing features and expand the library of available anonymization techniques.

4.2 Access Policies Management Component

4.2.1 Overview

The Access Policy Management Component serves as the policer of data access in the entire DIGITISE ecosystem. It is the fundamental service used to create, manage, and enforce the security, compliance, and governance policies that dictate how users and applications interact with data. The component and policies will allow for all data access requests to always be audited with respect to design defined rules. Therefore, the component will provide a single control point for a more transparency and trusted environment with more detailed controls of data permission.

4.2.2 Delivered Functionality

This section outlines the delivered functionalities of the Access Policy Management Component as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides a complete, end-to-end workflow for policy management, from creation to enforcement. The list of the delivered features is provided:

- Access Policy Configuration (APM-01): A comprehensive configuration interface has been delivered, enabling data providers to define and customize the rules that govern data access based on a set of predefined conditions.
- Access Policy Store (APM-02): A secure and reliable Access Policy Store has been implemented, ensuring that all configured policies are stored consistently and can be managed effectively throughout their lifecycle.
- Access Policy Distribution (APM-03): The policy distribution mechanism has been realized (alpha version), allowing the dynamic enforcement across the DIGITISE ecosystem to dynamically retrieve and apply the correct policies.

The component's scalable and secure architecture, as realized in this release, establishes a framework for integrating more complex policy models and enforcement mechanisms in the future.

4.2.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While the current framework supports robust rule-based policies, the initial release focuses on a core set of attributes and conditions. The incorporation of more advanced policy model with additional parameters will be considered for the 2nd version.
- Access policies are now defined as static and will require continuous review and updates. The component will be enhanced to enable on the fly updates on the access policies definition to easily modify and distribute new policy updates.

These constraints define the scope of functionalities available in the current version of DIGITISE. In the next release, the aim is to enhance the existing features, by expanding the list of parameters for the access policies as well as the logging for the access policies distribution.

4.3 Identity Provider Component

4.3.1 Overview

The Identity Management Component provides the framework for securing users authentication to enhance privacy over the interactivity within the DIGITISE ecosystem. The core function of the Identity Management is to manage the entire lifecycle of digital identities ensuring all users, applications, and data resources can be reliably authenticated and authorized. The component supports strong user authentication through many authentication methods (including SSO or Single Sign-On) which improves user experience by providing easier access to multiple services with a single set of credentials, creating a single trusted space for data management and sharing.

4.3.2 Delivered Functionality

This section outlines the delivered functionalities of the Identity Management Component as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides the essential tools for managing user identities and monitoring their activity, which is fundamental to the platform's security. The list of the delivered features is provided:

- User profile management (UM-01): The system provides comprehensive user profiles that include attributes and roles per user, enabling access to resources based on individual settings.
- User logs management (UM-02): The component tracks all authentication events and access patterns. This functionality supports compliance with data protection

regulations and organizational security policies, reinforcing trust among all stakeholders.

The Identity Management Component establishes a solid foundation for integrating more advanced identity and access management features in the DIGITISE project.

4.3.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- The first release provides a robust authentication mechanism considering specific roles and attributes during the configuration phase. Additional roles/attributes will be considered for the final release
- User identity and role management is an ongoing process that requires continuous synchronization with the practices of participating organizations. As the project evolves, the Identity Management Component will be considered to support SSO functionality for the different applications of the project.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for further enhancing the existing features, strengthening authentication options and improving identity lifecycle management.

4.4 Technology Stack and development details

Following the detailed presentation of the different components that consist of the Data Privacy and Security Layer of DIGITISE, an overview of the different technologies and frameworks considered for the delivery of the services is presented below.

Library	Version	License
NodeJS	18	MIT
Flask	-	MIT
Vue.js	2.7	MIT
TailwindCSS	2	MIT
Keycloak	-	LGPL-3.0
Pandas	1.4.4	New BSD
NumPy	1.26	BSD
Kafka	-	Apache License 2.0
Zookeeper	-	Apache License 2.0
PostgreSQL	-	PostgreSQL License (similar to BSD/MIT)
Redis	-	3-clause BSD License
MinIO	-	Apache License 2.0
MongoDB	-	Apache License 2.0
Vault	-	Mozilla Public License 2.0
SpiceDB	-	Apache License 2.0

TABLE 5 DIGITISE PRIVACY AND SECURITY TECHNOLOGIES

This section provides a high-level overview of the key technologies and frameworks used for the implementation of the Data Privacy and Security Layer.

- The ID Provider service utilizes Keycloak as an open-source Identity and Access Management solution responsible for authentication features such as user registration, Single Sign-On (SSO), and token generation (e.g., JWT). As a result of working with sensitive credentials and other secured things in the system, the infrastructure uses Vault for secure secret management.
- Access Policies Definition and Enforcement component role is two-fold. The policy definition interface, where administrators create rules, is built using a frontend stack of Vue.js and TailwindCSS, supported by a backend service in NodeJS or Flask. The defined policies themselves are stored in a structured database like PostgreSQL. For high-speed policy enforcement, the system utilizes SpiceDB as a specialized permissions database to evaluate access requests dynamically. To ensure low-latency checks at runtime, policies are also cached in Redis.
- The Data Anonymization component utilizes a Python-based data science stack to ingest datasets and perform operations with privacy-preserving techniques. For core data wrangling operations—choosing fields and performing statistical transformations—we employ Pandas and NumPy. Front end configuration is based on Vue.js and TailwindCSS.
- For ingestion, storage, and operational data processing following data anonymization (which continues the data governance layer's activities) via unstructured data large data files (a data lake), MinIO provides S3-compatible object storage, and for flexible semi-structured data we use MongoDB as the NoSQL document store.
- Any logging and communication among the different services is performed using Kafka brokers and Zookeeper as the coordination service as mentioned also in previous section.

Along with the presentation of the different open-source technologies considered for the delivery of the DIGITISE Privacy and Security layer, indicative screenshots from the 1st version of the application are provided to showcase the development progress on the work performed during the reporting period.

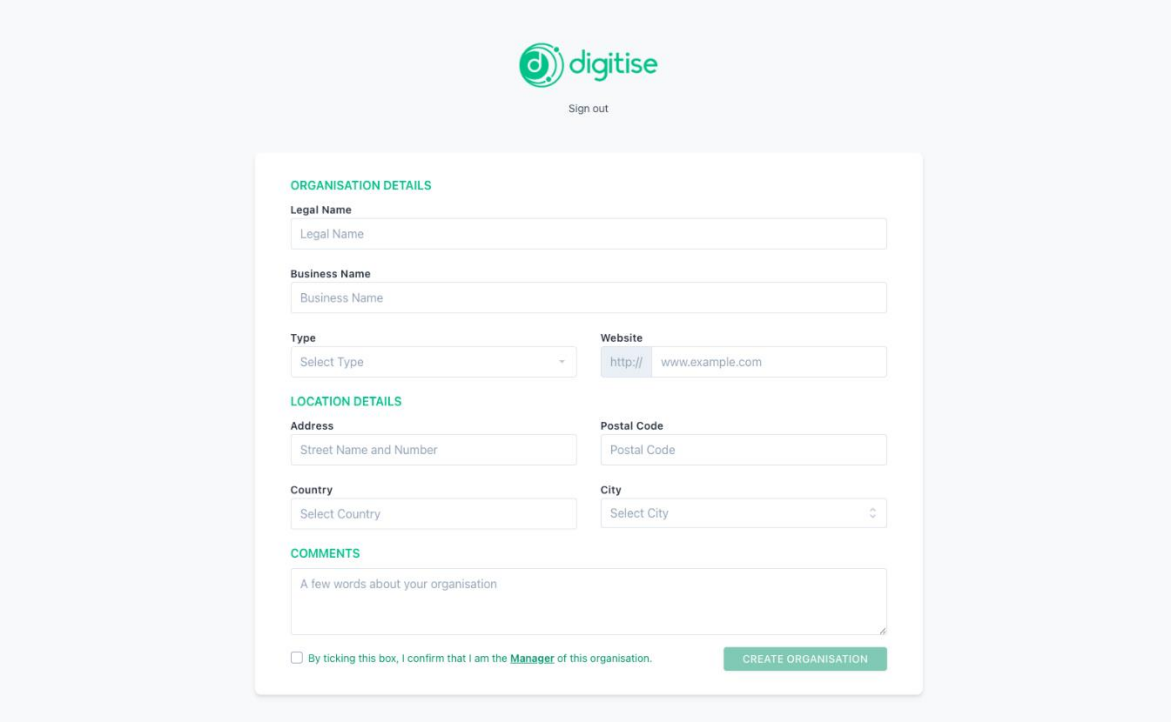

FIGURE 4 IDENTITY PROVIDER OVERVIEW

5 Information Sharing and Data Marketplace Environment

As stated in D2.2, the Data Sharing and Marketplace Layer focuses on facilitating access to data through advanced querying capabilities and marketplace functionalities. This layer enables users to search, discover, and retrieve relevant data assets, while also encouraging data sharing and reuse. Additionally, it supports the promotion of consumer participation in data-driven economic models, thus enabling the monetization or value exchange of data under well-defined and governed scenarios. The details of the different components (data marketplace, data exploration, data retrieval) as defined in D2.2 are presented below.

5.1 Data Marketplace Component

5.1.1 Overview

The Data Marketplace Component is the interaction point for any Marketplace or people to come together for data exchange within the DIGITISE ecosystem. It creates an online marketplace or digital catalogue as a comprehensive digital catalog where data providers can publish, describe, and monetize datasets. Data providers can richly describe their assets with metadata, including provenance, quality and structure information, and create flexible pricing and access policies. Data consumers can leverage discovery tools integrated in Marketplace for searching, filtering, and evaluating datasets to fulfill their specific needs. Overall, the Marketplaces functionality provides clear interaction and negotiation in developing a data economy, creating local and regional space for collaboration and innovation across the platform.

5.1.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Marketplace Component as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of the three core features provides a complete, end-to-end workflow for the data-sharing lifecycle, from initial asset discovery to contractual agreement and final settlement. The list of the delivered features is provided with focus on the features defined in D2.2:

- Data Marketplace Overview (DM-01): A discovery interface allowing data consumers to conduct effective searching and filtering of categorized data assets to find relevant datasets that meet their needs has been delivered.
- Contracts Management (DM-02): A fundamental Smart Contract Management capability has been delivered, where stakeholders, including Requesters, Approvers, and Publishers, can draft, negotiate, accept or reject, extend and amend contract terms all in one platform.
- Contracts Settlement (DM-03): A settlement engine has been integrated to ensure that signed smart contracts are effectively activated.

The component's robust and feature-rich architecture, as realized in this release, establishes a solid foundation for introducing more advanced e-commerce and community features in the future.

5.1.3 Considerations, Assumptions, and Constraints

In the initial release, constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While the first release provides a complete transactional framework, it focuses on basic discovery and contracting. Future iterations will introduce more advanced filtering for datasets search in the marketplace. In addition, the contracting process will be enhanced with additional attributes needed to define the smart contracts
- The value and utility of the Data Marketplace are directly dependent on the continuous contribution of data assets from providers. Towards this direction, the active enrolment during the alpha testing activities will enable us to review the overall functionality envisioned by the marketplace.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for users. In the next release, the aim is to enhance the user experience by building out these advanced marketplace features as stated above.

5.2 Data Exploration Component

5.2.1 Overview

The Data Exploration Component provides an interface for users to explore and comprehend data assets that are made available to them. This is an important first step after user acquisition, as it provides a set of views that can allow for first stage analysis and insight discovery. The component allows users to explore the characteristics, structure, and quality of a dataset using powerful search and visualization. Furthermore, the component helps users find specific information quickly, which will help close the distance between raw data and decisions, ensuring that further processing and validation can be effective and efficient.

5.2.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Exploration Component as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides users with the essential toolkit needed to inspect and analyze their acquired data assets effectively. The list of the delivered features is provided:

- Data Search (DE-01): A search functionality has been delivered, allowing users to query their acquired data assets based on keywords, metadata, and other filters to quickly find relevant information within their data pool.

- Data Visual Exploration (DE-02): The component provides a range of graphical representations (focus on hierarchical graph) to analyze and interpret data assets, enhancing the user's understanding of the data's characteristics and facilitating insight discovery.

The component's intuitive and interactive design, as realized in this release, establishes a solid foundation for building a more comprehensive exploration tool in the next phase.

5.2.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- While the first release provides essential search and visualization tools, it is focused on preliminary data exploration. The incorporation of more advanced search functions will be considered for future releases.
- The effectiveness of the data visualizations is dependent on the structure and type of the underlying data asset. As the variety of datasets in the marketplace grows, then the full-scale testing of the exploration engine will be performed to ensure an efficient operation of the overall engine.

Overall, in the next release the aim is to enhance the capabilities of the component, transforming it from an exploration tool into a more powerful preliminary analysis environment.

5.3 Data Retrieval Component

5.3.1 Overview

The Data Retrieval Service complements the Data Exploration engine in a similar way, and provides the means by which data can be accessed across the DIGITISE ecosystem, strictly according to the terms of the applicable contractual arrangement, while providing a consistent manner of retrieving data across the ecosystem. Access to data is through a clearly defined and secured set of API endpoints for requesting and consuming data through external applications. The retrieval part of the service is built on a powerful query configuration that allows users to define a range of needs by identifying sources, filtering and selecting output. Security is baked into the service, operating robustly in terms of authentication and authorization for every request in accordance to manage privacy and integrity of data. Finally, it should be noted that the application and the back-end give flexibility and a user-friendly interface and each user can configure the retrieval conditionally for their project.

5.3.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Retrieval Service as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides a

complete workflow for users to define precisely what data they need and receive it in the appropriate format. The list of core, delivered features is provided below:

- **Query Configuration (DR-01):** The service provides the functionality to define and customize data queries based on user requirements. This includes selecting data assets, applying attribute filters, setting parameters, and specifying output details to ensure efficient and relevant data retrieval.
- **Data Sharing (DR-02):** A core function to handle the retrieval and delivery of data "slices" has been delivered. The service ensures that retrieved data is formatted appropriately and aligns with the user-defined preferences established during the query configuration phase.

The component's secure and flexible API architecture, as realized in this release, establishes a foundation for supporting more advanced data retrieval patterns and optimizations in the future.

5.3.3 Considerations, Assumptions, and Constraints

The initial release has some limitations that will impact how effectively the DIGITISE Data Space can be deployed.

- The first release provides a mature query and retrieval capability but any enhancements such as intelligent caching, server-side processing, and effective pagination functionality for accessing large data records are future optimizations that have been identified.
- The performance of the Data Retrieval Service is reliant upon the underlying Data Storage layer, and the complexity of user queries. As the system expands, will monitor query patterns in the system to better understand any performance optimization opportunities.

Any constraints will determine what functionality has been made available in the current generation of DIGITISE and will be an important consideration to plan new features to assist with more intensive data access use cases on a mass quantity scale.

5.4 Technology Stack and development details

Following the detailed presentation of the different components that consist of the Data Information Sharing and Data Marketplace Layer of DIGITISE, an overview of the different technologies and frameworks considered for the delivery of the services is presented below.

Library	Version	License
NodeJS	18	MIT
Vue.js	2.7	MIT
TailwindCSS	2	MIT
Kafka	-	Apache License 2.0
Zookeeper	-	Apache License 2.0

Library	Version	License
PostgreSQL	-	PostgreSQL License (similar to BSD/MIT)
Redis	-	3-clause BSD License
MongoDB	-	Apache License 2.0
Vault	-	Mozilla Public License 2.0

TABLE 6 DIGITISE INFORMATION SHARING AND DATA MARKETPLACE TECHNOLOGIES

This section provides a high-level overview of the key technologies and frameworks used for the implementation of the Information Sharing and Data Marketplace Layer.

- Data Marketplace and Contracts combines traditional web technologies. A user-based marketplace built with Vue.js and TailwindCSS, supported by NodeJS for the backend. This part uses a relational database like PostgreSQL to store detailed metadata about data assets, pricing models, and negotiation history. Sensitive credentials are securely managed by Vault.
- The Data Exploration Component provides users front-end built using Vue.js and TailwindCSS. This user interface allows for the visualization and preliminary analysis of acquired datasets. The backend logic that processes data for visualization and serves it to the front end is developed in NodeJS.
- The Data Retrieval Service is primarily a backend API built with NodeJS. It manages secure, on-demand access to data slices based on the rules defined in the smart contracts. It uses PostgreSQL to store query configurations and leverages Redis for caching frequently accessed data to ensure high-performance delivery.
- Asynchronous event notifications between the marketplace and retrieval services, such as contract updates, can be handled via a messaging system Kafka with coordination from Zookeeper as in previous sections.

Along with the presentation of the different open-source technologies considered for the delivery of the DIGITISE Information Sharing and Data Marketplace layer, indicative screenshots from the 1st version of the application are provided to showcase the development progress on the work performed during the reporting period.

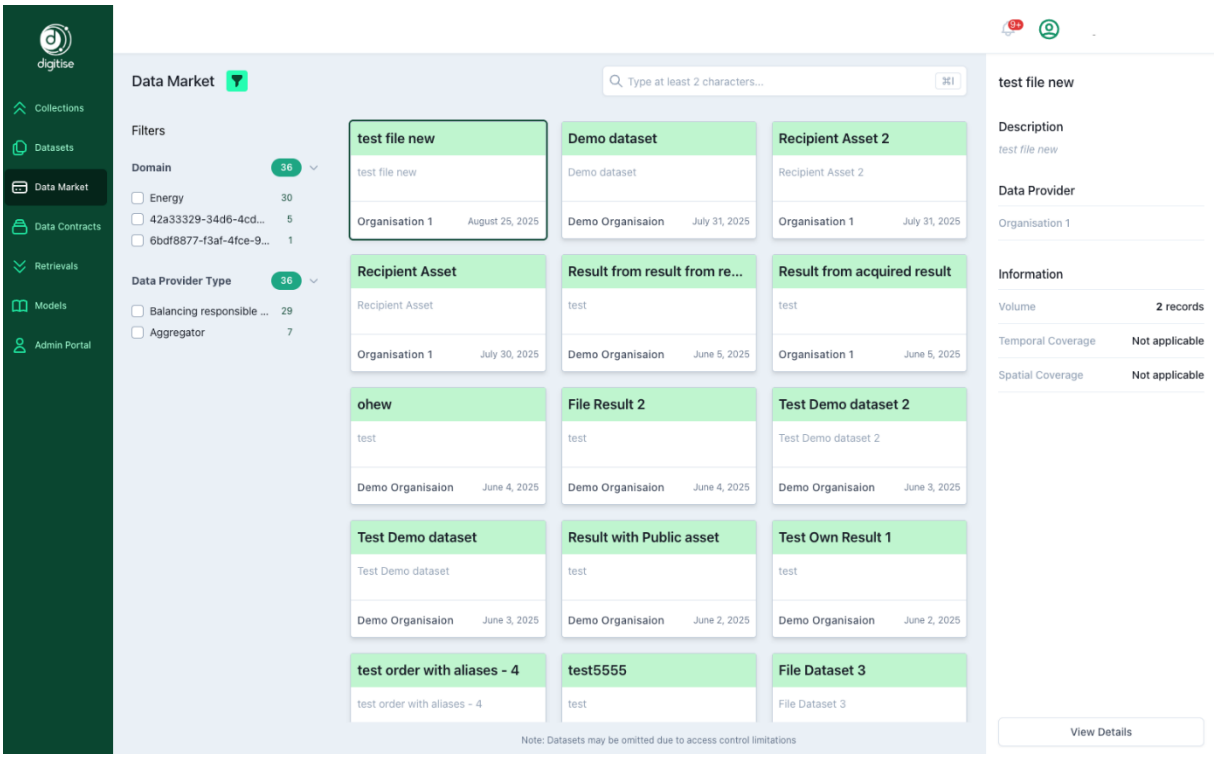

FIGURE 5 DATA MARKETPLACE OVERVIEW

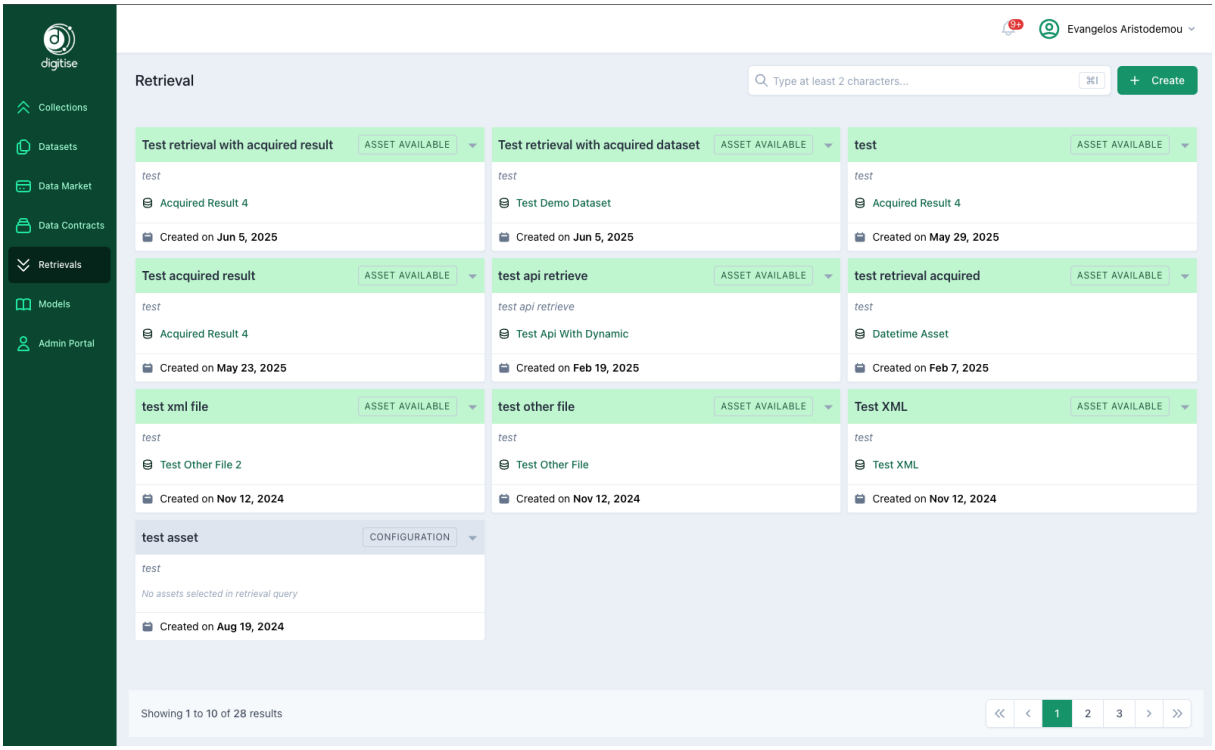

FIGURE 6 DATA RETRIEVALS OVERVIEW

6 DIGITISE Data Analytics Models

As stated in the DoA but also D2.2, the core of this task is to provide a suite of advanced AI analytics models designed to translate raw household data into actionable intelligence. The analytics are structured into three interconnected layers:

- Personal Analytics: This initial layer focuses on creating a deep, individualized understanding of the household, establishing baseline profiles for consumer energy behavior, occupant presence, and comfort preferences. By creating this analysis, we may enable provision of personalized energy and non-energy services (such as security or health monitoring) as the scope of DIGITISE.
- Short-Term Energy Forecasting: As to produce accurate forecasts of electricity demand and on-site generation (e.g., solar) from 1 hour up to one day in advance.
- DER Flexibility Analytics: As to profile, forecast, and characterize the energy flexibility available from the Distributed Energy Resources (DERs) in a home (e.g., HVAC systems, water heaters, EV chargers).

By taking into consideration the project needs as reported in D2.2, the non-exhaustive list of models to be delivered in the project is presented below following the aforementioned taxonomy.

Model Name	Description	Category	Comment
Short-Term Demand Forecasting	Provides hour-to-day-ahead predictions of household energy demand, ensuring accurate load planning.	Short-Term Energy Forecasting	Available in the 1st version
Short-Term Generation Forecasting	Predicts renewable energy generation (e.g., solar, wind) in the short term to optimize self-consumption and grid interactions.	Short-Term Energy Forecasting	Available in the 1st version
Consumer Energy Behavior Analytics	Analyzes household energy consumption patterns to extract insights on how consumers use energy, helping to create personalized energy-saving strategies.	Personal Analytics	Partially available in the 1st version
Occupancy Profile Analytics	Monitors occupancy data to understand when and how different spaces are used, improving energy efficiency and automation.	Personal Analytics	Partially available in the 1st version
Comfort Preference Analytics	Assesses user preferences for heating, cooling, etc. optimize comfort while minimizing energy waste.	Personal Analytics	Available in the 1st version
Ambient Condition Analytics	Tracks environmental parameters such as temperature, humidity, and air quality to enhance energy management and indoor comfort.	Personal Analytics	Available in the 1st version

Short-Term Demand Forecasting/Device Level	Provides hour-to-day-ahead predictions of device level energy demand, ensuring accurate device management	Short-Term Energy Forecasting	Available in the 1st version
Context-Aware Flexibility Profiling	Integrates real-time data, device characteristics, environmental conditions, and occupant behavior to dynamically assess available flexibility.	DER Flexibility Analytics	Available in the 1st version
Batteries Flexibility Profiling	Analyzes batteries as flex assets to determine their flexibility potential	DER Flexibility Analytics	Available in the 2nd version

TABLE 7 LIST OF ANALYTICS IN DIGITISE PROJECT

Note I: The analysis of context conditions was partially addressed in this version. Updates on the models for personal analytics will be performed in this second period. Also, the analysis of batteries as flexibility assets to determine their flexibility potential will be included in the second version of the project. This is due to the need to have accurate demand and generation profiling results as a prerequisite for this analysis. While the results from demand/generation profiling were made available during this period, further evaluation of these models is needed.

Note II: To further enhance the analytical capabilities of the DIGITISE ecosystem, a roadmap for future model development is under consideration in consortium in line with data availability from the project's pilot activities. A key area of focus will be the development of detailed models for profiling Electric Vehicle (EV) charging points and estimating their demand-side flexibility. The feasibility and scope of this work will be evaluated following the successful acquisition and validation of relevant datasets from the demonstration sites. Additionally, the project plans to develop models for profiling the generation flexibility of Distributed Energy Resources (DERs). The development of these models is conditional upon receiving detailed specifications regarding the technical controllability of the generation assets, which will enable an accurate assessment of their potential contribution to the energy system.

As also stated in D2.2, the overall technical approach is to provide a complete workflow for energy analytics through three layers: (a) AI Configuration to build, test, and refine machine learning models. (b) AI Models Storage to set the environment for models' storage and (c) AI Execution to runs them efficiently and report the results. On the basis of this conceptual approach, the details about the implementation of the different models considered in the project are listed below.

6.1 Short-Term Generation Forecasting

6.1.1 Overview and Business Scope

Short-term generation forecasting plays a crucial role in the DIGITISE project as a key enabler for local energy optimization and demand-side flexibility orchestration. This forecasting component specifically targets the prediction of residential solar photovoltaic (PV) system output up to 24 hours in advance with hourly resolution. The forecasting model is designed to account for the variability in solar generation by incorporating both historical production data and near-real-time contextual variables such as solar irradiance, ambient temperature, humidity, and wind speed. Additionally, PV system characteristics (e.g., orientation, tilt, capacity) and geographic information are used to tailor the forecast to the physical reality of each installation.

This component is deeply integrated into the DIGITISE pool of analytics, aligning with the project's broader vision of context-aware and explainable energy intelligence. Generation forecasts are not treated as isolated outputs but serve as a foundational input to baseline definition, flexibility estimation, and transparency in energy transactions, especially when validated against actions from control agents.

6.1.2 Technological Implementation and Deployment

The technical implementation follows a progressive modelling approach that begins with a Gradient Boosting Machine (GBM) model, specifically the LightGBM model in the Darts library. This model is selected for its efficiency and performance in handling structured time series data, especially in the presence of noisy or missing values.
The training dataset includes:

- Historical PV power output data.
- Weather context: solar irradiance, ambient temperature, humidity, and wind speed.
- System metadata: PV panel tilt, orientation, rated efficiency.
- Temporal features: hour of day, day of week, seasonal indicators.
- Proxy features.

To ensure robustness, a fallback hybrid deep learning architecture is foreseen. This model combines convolutional neural networks (CNNs), which detect local generation pattern shifts, with long short-term memory (LSTM) layers that capture temporal dependencies in solar production. This layered design ensures high adaptability across different user and asset profiles.
The model will be trained using an expanding-window strategy, gradually incorporating new operational data while preserving sequence integrity. Forecasts will be generated and updated on an hourly basis. These outputs will be fed directly into DIGITISE's business applications and user interfaces, where they support both planning and real-time energy

decisions. Generation forecasts also contribute directly to flexibility validation. By establishing accurate, up-to-date generation baselines, they help assess the feasibility and effectiveness of control actions, especially those involving demand-supply balancing and prosumer engagement.

6.2 Short-Term Demand Forecasting

6.2.1 Overview and Business Scope

Short-term consumption forecasting is an essential capability in DIGITISE, directly supporting household-level energy optimization, proactive control, and the reliable activation of demand-side flexibility. The goal is to estimate residential electricity consumption over the next 24 hours with hourly granularity. These predictions form the basis for planning energy dispatch, triggering load shifting strategies, and identifying flexibility potential particularly where sub-metering is limited or absent. Developed by CIRCE, the consumption forecasting solution extends beyond static historical trend analysis by embedding behavioral, environmental, and contextual dynamics. The model considers thermal conditions, occupancy proxies, and indirectly user routines, enabling predictions that align with real-time building operation and lifestyle patterns.

Aligned with DIGITISE's AI analytics framework, this forecasting tool integrates seamlessly into flexibility profiling and orchestration activities. By delivering near-real-time demand predictions, it supports transparent baseline definition and enhances user-led control strategies.

6.2.2 Technological Implementation and Deployment

The initial consumption forecasting implementation is based on LightGBM, chosen for its ability to handle large, structured datasets and its effectiveness in capturing non-linear relationships. The model is trained using a comprehensive set of historical and contextual variables, including:

- Consumption Data: total household electricity usage.
- Environmental Inputs: indoor and outdoor temperatures
- Temporal Features: hour of day, weekday/weekend flags, holiday indicators, and seasonal trends
- Behavioral Proxy: inferred occupancy.

In households lacking fine-grained sensor data, proxy variables derived from time-based features are used to approximate user activity and comfort routines. Model training follows an expanding-window approach, which allows the model to continuously adapt to new data while maintaining chronological integrity. This ensures the model evolves in line with household behavioral changes and seasonal load variations.

Moreover, when a need arises, the forecasting system may be enhanced with hybrid deep learning approaches to improve sensitivity to complex temporal consumption patterns. Future extensions could integrate LSTM architectures, particularly in cases where fine-

grained consumption data is available and behavioral variability is high. The forecasts are updated every hour and fed into the DIGITISE orchestration and flexibility estimation layers. These predictions support short-term decision-making for load dispatch and also provide input to flexibility quantification—by comparing predicted "business-as-usual" demand with potential demand-reduction scenarios.

Multiple versions of the consumption forecasting model are under development to support varying data granularities and market contexts across DIGITISE pilots. These versions are designed to be robust to different regulatory, technical, and spatial conditions, ensuring scalability and transferability across EU member states.

6.3 Consumer Energy Behavior Analytics

6.3.1 Overview and Business Scope

Consumer Energy Behavior Analytics is a foundational element of the DIGITISE platform, designed to extract, model, and interpret patterns in residential energy use beyond mere consumption values. It focuses on understanding how, when, and why energy is used in the household, thus supporting behavior-aware control and personalized recommendations. This component transforms raw consumption data into actionable behavioral insights. By analyzing the temporal and contextual characteristics of energy use, it identifies routines, anomalies, and emergent patterns that reflect lifestyle dynamics. Household characterizations are extracted e.g., energy-intensive, efficiency-driven that on one hand inform the prosumers about their behaviours, increasing their energy literacy. On the other hand, it allows such analytics to be used for tailored recommendations that assist in lowering the energy footprint of the households.

There exists a vast line of works that focuses on extracting behavioural traits of households, finding application in multiple areas, e.g., in load prediction [58] and demand response programs [59]. The state of the art revolves around employing machine learning techniques with a primary goal of clustering consumers based on their behavioural profiles. To that end, multiple architectures have been explored, implemented, and tested, showcasing the different trade-offs between them e.g., in terms of accuracy and supported features. For example, by using DBSCAN for consistency analysis yields the following advantages with respect to other techniques: there is no need to define a priori the number of clusters in the data, and it supports the notion of outliers [60]. Specifically for electricity consumption now, it has been shown that arising patterns will repeat with noise [61]. Therefore, DBSCAN is an ideal clustering technique to identify outliers from a set of daily consumption profiles. Notably, having fewer outliers in the result, indicates a better consistency in the system overall. Another fitting candidate for behavioural clustering in energy related applications is K-means. The authors of [62] utilized K-means to identify distinct patterns within aggregated load profiles for each day of the week to

support network operations and to examine typical consumption patterns and individual daily profiles of households for cost-effective targeting of the most appropriate set of households for demand response schemes. More recently, in [63], the authors employed unsupervised classification techniques or real historical data. to extract groupings of consumers on the basis of how similar their patterns are in electricity consumption. Last, a combination of the two methods might yield the best results, as demonstrated in [64].

6.3.2 Technological Implementation and details

To realize the behavioural analytics engine, a multi-layered architecture is employed. It combines statistical profiling, unsupervised learning, and time-series pattern recognition. Utilizing consumption data from the households alongside other potential sources (e.g., temporality-dependent data, weather conditions, and indoor temperatures) the engine outputs a prediction on the expected behaviour and later identifies deviations or irregularities. For household segmentation, clustering algorithms are utilized (e.g., K-MEANS, DBSCAN) and deep neural networks are employed for the model architecture that extracts desired features and predictions. Notably, the engine is able to provide users with dynamic profiles based on predefined time periods (e.g., week, month, year).

6.4 Occupancy Profile Analytics

6.4.1 Overview and Business Scope

Occupancy Profile Analytics estimates the presence profile of the household, crucially, without the need for intrusive sensors. This allows for behaviour-aware control strategies, leading to better and more efficient energy management of the household. The key objective is to generate probabilistic or binary information regarding occupancy with high-enough temporal resolution, offering insights into household usage patterns that influence energy demand and flexibility availability. Notably, this analytics layer enables the derivation of dynamic occupancy maps that evolve with household routines and can adapt to seasonal, behavioral, or operational shifts.

In the literature there is a strict distinction between intrusive and non-intrusive occupancy monitoring. As we are focusing on extracting occupancy patterns from smart meter and electricity consumption data (which is non-intrusive) a plethora of techniques are suitable to realizing our goal. For example, one might analyze a time use survey (TUS), aiming to report data on how people spend their time. Its objective is to identify, classify, and quantify the main types of activity in which people engage during a definitive time-period [65]. TUSs are versatile and can be combined with multiple methodologies to extract meaningful occupancy profiles, at household levels. More specifically, in [66] a bottom-up modeling approach was used, together with a set of calibration methodologies, to predict

residential building occupants' time-dependent activities for use in a dynamic building simulation. However, in [67] another approach was followed, one that revolves around extracting typical occupancy profiles via statistical methods. In fact, the authors conducted a clustering analysis using TUS achieved occupancy profiles for archetypal building models. More recently, in [68] the authors propose a multi-stage process that incorporates an occupancy detection algorithm whose purpose is to extract occupancy profiles. They utilize a threshold-based method, which logs information related to plugs and lighting electrical consumption. These are given binary values at any point, based on specific thresholds (which one can dynamically configure). Essentially, the output of this process is a time series of binary hourly values, indicating the occupancy of each individual space/household. Finally, an approach where multi-modal analysis was performed using non-intrusive household data has been proposed in [69]. Using traditional techniques alongside machine learning models, the authors extract occupancy profiles. More specifically, to achieve their goals they rely on clustering techniques alongside linear regression modeling.

6.4.2 Technological Implementation and details

The implementation uses both heuristic and learning-based methods. Using data from available sensors (e.g., electricity consumption, indoor temperature, smart thermostat settings) and external conditions (e.g., outdoor temperature, daylight hours) various features are extracted. It identifies temporal proxies (e.g., peak and quiet hours), sudden load changes, and heating and cooling system activation periods. For basic presence detection it relies on rule-based classifiers and Bayesian inference models for probabilistic occupancy estimation. Additionally, supervised models assist in generalizing occupancy patterns across broader household communities. Last, occupancy profiles are dynamic (e.g., updated every 15 minutes or hourly).

6.5 Comfort Preference Analytics

6.5.1 Overview and Business Scope

Comfort Preference Analytics is a user-centric module designed to infer and model individual or household thermal comfort preferences based on system usage patterns, indoor environmental conditions, and user control actions. This enables personalized and adaptive control strategies (e.g., for HVAC systems) that align energy efficiency goals with occupant satisfaction. This module supports adaptive control loops, includes personalization elements, and utilizes feedback mechanisms to promote behavioural change. In any case, the households' comfort zone is at the center of any suggestions or automated actions. These comfort zones are dynamically set by the users as fixed setpoints but also adaptable based on temporal elements (e.g., seasonality).

Multiple solutions have been proposed to derive human comfort levels from raw sensor data [70]. Depending on the typing of the available data and the application scenario different analytical methods and machine learning algorithms have been utilized. For example, in [71] the authors construct a physics-based autoregressive moving average model for room temperature in office buildings. Using thermodynamic equations, they determine the structure and order of a linear regression model, which can predict the room temperature over several weeks into the future. Notably, one main, motivating observation of this work was that while thermodynamics models indeed provide the most comprehensive description of the thermal processes in buildings and accurate estimations of various system outputs, at the same time, they consist of a considerable number of parameters, have inherent high complexity, and uncertainty in thermal properties of structural elements make it quite a challenge to obtain an accurate modeling process. In fact, more recent works rely on learning methods to overcome such challenges. The authors of [72] tested a variety of different architectures (Classification Tree, Gaussian Process Classification, Gradient Boosting Method, Kernel Support Vector Machine, Random Forest, and Regularized Logistic Regression) and reported their respective performance. Importantly, this study was aimed at individual occupant comfort levels, despite being an inherently subjective phenomenon which can display large differences among individuals. One of the main derivatives is the tradeoff between computational power and training required and achieved accuracy for the different architectures considered. Last, similarly to above, the authors of [73] employ four different learning algorithms (Support Vector Machine, Artificial Neural Network, Random Forest, and Extreme Gradient Boosting) in their design and compare the accuracies. The main takeaway is that there is no clear winner, and depending on the target metric to be evaluated different techniques outperform others (e.g.., Random Forests might provide higher accuracies in general, however, Artificial Neural Networks have superior performance when predicting considerably low occupant-predicted mean vote values.

6.5.2 Technological Implementation and details

For implementing this module, a combination of time-series modeling, supervised learning, and statistical inference is used. Sensed temperature data from inside the household in combination with thermostat setpoints and occupancy patterns are used as input. The supported features revolve around the estimation and establishment of preferred temperature band during various occupancy periods, adjustment based on seasonal patterns, and automation via readjustment of the frequency and deviation aspects of the comfort profile and based on the available actions. Regarding modeling, supervised learning is employed for comfort band likelihoods, and time-based comfort levels are

calculated. These analytics are computed daily and incorporated into the orchestration and control layers.

Dataset

The analysis uses an open-source dataset available at EU Data Portal. The dataset contains measurements from three building blocks (A, B, and C). For this study, only data from block A was considered.

Data Preparation

Since most of the data had outliers and missing points, these were filled using a forward-fill method. The cleaned data was then resampled into hourly averages. The HVAC dataset includes user temperature setpoints, the operating mode of the system, and on/off status. Similar resampling and cleaning were performed. Additional features were created, such as the time of day (morning, afternoon, and night) and the season of the year (summer, autumn, and winter). The weather dataset contains outdoor temperature, humidity, solar radiation, precipitation, and other variables. Since each variable included minimum, maximum, and mean values, We used correlation analysis to remove redundancy. Only the mean values were kept for the analysis. All three datasets were resampled to one-hour intervals and merged into a single dataset for modelling.

Comfort Preference Modelling

Initial analysis shows that setpoint was maintained between 22–23 °C in all seasons. Indoor temps raised significantly in spring and summer, reaching ~32 °C in summer, well above the 22–23 °C setpoint. This indicates either limited HVAC usage/capacity, or that HVAC systems were not able to fully maintain the desired comfort. Our initial model is set to forecast indoor conditions and be able to recommend changes that maintain indoor conditions while minimising HVAC ON time / unnecessary cooling/heating.

To this end, we explored two popular forecasting models: RandomForestRegressor, which is well-suited for capturing nonlinear relationships and can incorporate external features, and SARIMAX, a time-series model that accounts for seasonality while also using exogenous variables. Since the data was timeseries based, we used time-ordered split (with 0.7, 0.2 and 0.1 for training, validation and testing sets respectively), the same features, and report (MAE and RMSE) metrics were used. Initial result showed that SARIMAX was able to outperform RandomForestRegressor.

The goal is to give users easy recommendations. For example, if the forecast shows the indoor temperature will reach 26 °C at 3 pm, the system suggests lowering the setpoint to 23 °C in advance to pre-cool, save energy, and stay comfortable. If the forecast shows the

temperature will remain steady at 22.2 °C overnight, the system suggests keeping the current setting so the HVAC can stay off and save energy.

Next steps will focus on refining model accuracy through parameter tuning and testing. The other stage will involve integrating forecasts into a control framework for setpoint recommendations and validating the approach on other building blocks to assess generalizability.

6.6 Ambient Condition Analytics

6.6.1 Overview and Business Scope

Ambient Condition Analytics provides a comprehensive, contextual understanding of the physical environment within and around residential buildings. This module captures and analyzes environmental parameters such as indoor temperature, humidity, and outdoor weather conditions, which are critical drivers of thermal comfort and energy efficiency. By continuously monitoring and interpreting these variables, it can better anticipate energy demands, optimize operations, and refine behavioral models with greater environmental accuracy, leading to suggestions/decisions towards better living conditions for the household.

Various strategies have been proposed in the literature to analyze and predict ambient conditions in households or business-purposed spaces for example. However, we observe a convergence in the more recent years, with technologies relying on machine learning seem performing better in terms of efficiency and accuracy than folklore techniques. In [74] the authors provide an extensive analysis of directions prior works have followed in the indoor air quality prediction, including their respective strengths and weaknesses. In terms of machine learning techniques, linear regression, decision trees, K-nearest neighbors, and support vector machines were considered to try and obtain, by comparison, the most reliable results. One interesting observation the authors make is that robust predictions often arise by taking the average across various models' outputs. Another approach was followed in [75], where the authors develop a prediction model for indoor air quality using region-based convolutional neural network (CNN) models with promising results. Notably, combining this approach with automated or assisted by recommendation control over various aspects of the household can significantly improve the living conditions of the occupants. Furthermore, various researchers have developed long short-term memory (LSTM) networks with similar results. For example, in [76] LSTMs were combined with generic algorithms, which are metaheuristic optimization algorithms. Fusing these techniques exploits the different strengths of each component. On one hand, LSTMs maintain the state of the memory even after a long time due to the presence of the memory cell in the architectural design, yielding significant advantages in the output

predictions. On the other hand, the parameter selection in LSTMs is challenging. However, metaheuristics genetic algorithms have been shown to be able to find the best parameters e.g., for the window size and the number of units in LSTMs. Additionally, the LSTM input sequences are of fixed length, and using genetic algorithms leads to a more flexible performance which is necessary when predicting pollution levels. Provably, this new approach for predicting next-day air quality is more efficient than determining the parameters manually. Last, a combination of CNN and LSTM architectures has been proposed in [77]. The main idea revolves around using the architecture's convolutional part as a feature extractor before the LSTM. Interestingly, while this fusion of CNNs with LSTMs proved a valuable modeling approach for this target application, it has certain limitations, mainly due to data variance as the authors observed.

6.6.2 Technological Implementation and details

This component is implemented through real-time data integration, sensor data processing, and predictive modeling. Regarding the sensor data inputs, it relies on indoor temperature, humidity, and CO_2 levels. It also utilizes other external data e.g., outdoor temperature, humidity, and wind speed. Last, it combines both the above with thermal comfort indices and calculates various outputs. Via this process, the difference between the indoor and outdoor temperature is calculated, and other environmental anomalies. This module relies on predictive models using LSTM and CNN architectures to achieve its functionality. Notably, such insights are updated periodically (e.g., every 15 minutes or hourly), depending on data availability.

Dataset

The analysis uses an open-source dataset available at EU Data Portal. The dataset contains measurements from three building blocks (A, B, and C). For this study, only data from block A was considered.

Data Preparation

Since most of the data had outliers and missing points, these were filled using a forward-fill method. The cleaned data was then resampled into hourly averages. The HVAC dataset includes user temperature setpoints, the operating mode of the system, and on/off status. Similar resampling and cleaning were performed. Additional features were created, such as the time of day (morning, afternoon, and night) and the season of the year (summer, autumn, and winter). The other dataset used for this, is the CO2 levels (measured in ppm). All three datasets were resampled to one-hour intervals and merged into a single dataset for modelling.

Implementation

We implement a two-layer pipeline: a multi-output forecaster and a multi-objective controller. Hourly HVAC logs, indoor sensors, and weather are aligned and enriched with lagged signals (indoor_temp_lag1, co2_lag1, setpoint_lag1, hvac_on_lag1), short rolling means (3-hour), and categorical context (Hour, Season, mode, device_type).
The forecasting model used XGBoost within a MultiOutputRegressor to jointly predict next-hour indoor temperature and CO_2 (indoor_temp_next, co2_next). Training was performed with an 80/20 time-ordered split, applying numeric standardization and categorical one-hot encoding. Model performance was assessed using MAE and RMSE. For comparison, a RandomForest-based multi-output model was also evaluated. Both models achieved strong predictive accuracy, with XGBoost showing slightly better performance for CO_2 forecasts, while RandomForest performed marginally better for indoor temperature. Overall, the results demonstrate that both approaches are suitable for short-term comfort and air quality forecasting.
While both RandomForest and XGBoost performed well, their differing strengths suggest potential for further exploration. We aim to explore more models and also define our multi-objective controller model for user comfort. The goal is to use these predictions as input for comfort model and be able to make suggests to improve user comfort while minimising energy consumption.

6.7 Short-term demand forecasting at device level

6.7.1 Overview and business scope

Demand forecasting in the building-centric energy domain holds an important role in demand response (DR) strategies definition, flexibility and further energy market participation, self-consumption optimization and control strategies extraction. When forecasting depth reaches the device level, the gained knowledge for future device consumption can be exploited as input for all the above tasks, but also for predictive maintenance and recommendation systems for energy efficiency or comfort preservation activities. In the scope of DIGITISE project, demand forecasting is focused on HVAC and DHW systems, since they comprise high energy demanding appliances from which data will be present from the different demo sites of the projects. HVAC and DHW units also hold a significant position in the DR control strategies and occupants' well-being and comfort maintenance, either explicitly by direct controlling of devices, or implicitly through recommendations of optimal device usage with proposed operation schedules and setpoints.

From the findings in recent literature, the most commonly used techniques for both DHW and HVAC short-term demand forecasting remain the deep learning (DL) algorithms like RNN (mainly LSTMs) and hybrid combinations of them (i.e. CNN-LSTM) but also SVR, ensemble and decision tree algorithms, regression algorithms, well-known time series and

stochastic models are also broadly studied with promising results. In the case of HVAC demand forecasting, authors of [48] explored a Support Vector Regressor in combination with optimization algorithms to improve forecasting results, whereas in [49] an ensemble of machine learning techniques have been tested, with XGBoost and Gradient Boosting Machine algorithms, accompanied by an exhaustive feature selection process, standing out with a good prediction accuracy. In [50], ANN and Random Forest (RF) algorithms were explored and applied on Spanish heating demand data, where ANN proved to continuously outperform the RF one. A number of different ML and DL algorithms were tested in [51] for forecasting an air handling unit's supply air temperature, that directly affects the unit's demand. Extra Tree, XGBoost, LSTM and CNN were compared, and the CNN+LSTM combination proved to provide the best results, since CNN could learn local features and reduce the number of parameters for making the LSTM training more efficient. Based on this advantage, authors of [52] proposed a bidirectional LSTM (biLSTM) with a CNN layer, but also combined with a CEEDMAN decomposition step at the beginning of the architecture, after which, the extracted signals along with exogenous variables are provided as input to the NN. In case of DHW demand forecasting, the authors of [53] opted for time series forecasting algorithms, namely exponential smoothing, basic structural and state-space models, where the results showed that state-space models and exponential smoothing outperformed the basic structural one, with the proposed solution to be the exponential smoothing as a simpler approach. Similarly, a wide list of timeseries forecasting models has been explored by authors of [54], among which was seasonal naïve model, exponential smoothing, ARIMA, seasonal decomposition method and combinations of them, which seemed to perform well on aggregated data of numerous households. An interesting alternative to the usual time series, machine learning and deep learning techniques was presented in [55]. A novel approach of continuously learning algorithms was proposed, for dealing with the catastrophic forgetting phenomenon of the commonly used deep learning algorithms, and the authors experimented with three models from the replay methods group of continual learning algorithms. An ANN was built with hyperparameter tuning in [56], where the goal was to compare the ability of accurate forecasts for either individual or aggregated data of multiple consumers, with the performance of the NN models to be adequate for the individual ones, but certainly better for a larger set of households included in the dataset. In the case of [57], two hybrid models were tested, an SARIMA time series forecaster combined with either an SVR or an LSTM, where in both cases the residuals of the SARIMA model are passed as input to the next step. In all tests that were performed, the SARIMA-LSTM hybrid model provided better results with less computation time needed.

6.7.2 Technological and implementation details

For the development of the two forecasting pipelines in the DIGITISE project, two different cases, for HVAC and DHW demand respectively were explored.

For the HVAC short-term (day-ahead) load prediction, a Long Short-Term Memory (LSTM) model was tested. Since HVAC usage has proved to have a strong reliability on previous measurements, an LSTM network was chosen for the initial trial of the forecasting method. The data used derived from an open dataset (Plegma)[2], where HVAC power consumption measurements (measured in Watts) with 10 sec granularity were aggregated in hourly intervals. The data preprocessing also included the following steps:

- Data scaling (min-max scaling)
- Creation of 24hour lag and 24hour log values (the latter used as the target values)
- Split of the dataset in train, validation and test sets, where two months were used for training and validation and 1 week for testing purposes

The rest of the training process was comprised of different tests on the depth of the network, the size of the units, kinds of additional layers (i.e. dropout ones), learning rate and more, in order to define the one with the best performance. As a next step in the implementation, is to integrate external variables, like internal and external temperature measurements, and also explore gradient boosting algorithms, like XGBoost and GBM. The training was performed with 24hour prediction horizon, but it can easily be converted to 15min day-ahead forecasting.

For the DHW day-ahead forecasting, a simple ANN was built and tested, as basis for future implementations and experimentation, since in many cases it provides promising results in time series forecasting. The data used for training derived from open datasets, Plegma and Almanac[3]. The Almanac data contained both water and power consumption data of the water heater, that allowed us to explore the usage patterns and get ensured that load measurements could also be used for forecasting purposes, since in the studied literature, only water consumption was used. For the different datasets the following preprocessing took place:

- Almanac data:
    - aggregation of original power consumption measurements (in Watts) from 1 minute to 15 minutes granularity
    - Feature engineering: the additional features for expressing the cyclical nature of the consumption where the hour, day of the week, month, weekend (or not), and lag measurements of 1, 2 and 24 hours before
    - Splitting of dataset in train, validation and test sets: a month of consumption data was set for training, 1 week for validation and another week for testing
    - Min-max scaling of the data sets

[2] https://www.nature.com/articles/s41597-024-03208-0
[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FIEOS4

- Creation of 24hour log values (in 15min granularity) to be used as target variables

- Plegma data:
  - Addition of other external variables since they were available in the dataset: internal and external temperature in Celsius
  - Interpolation, mainly for dealing with missing boiler consumption data
  - Aggregation of original power consumption measurements (in Watts) from 10 seconds to 15 minutes granularity
  - Feature engineering: the additional features for expressing the cyclical nature of the consumption where the hour, day of the week, month, weekend (or not), and lag measurements of 1, 2 and 24 hours before
  - Splitting of dataset in train, validation and test sets: 1.5 months of consumption data was set for training, 1 week for validation and another week for testing
  - Min-max scaling of the data sets
  - Creation of 24hour log values (in 15min granularity) to be used as target variables

In both dataset cases, a number of ANN networks were built and tested, and also a hyperparameter tuner was used for fine tuning of the different aspects that were tested, like the number of network hidden layers, units, optimizers, batch sizes and more. For improving the monitored performance scores, either a combination of the ANN with a timeseries forecasting method like SARIMA and exponential smoothing, or an LSTM network with a CNN layer will be considered for experimentation in the next version of the forecasting method.

6.8 Context-Aware Flexibility Profiling

6.8.1 Overview and Business Scope

Context-Aware Flexibility Profiling provides a new level of sophistication in how to characterize and utilize building energy flexibility. Instead of focused on a single static calculation, it treats demand-side flexibility as a convergent and intelligent assessment focused on the most impactful systems (HVAC, domestic hot water heaters) in buildings. The key concept is a building's capable of shifting energy usage is not set in stone and is influenced by many different layers of real-time data. Towards extracting accurate flexibility profiles, analyzing external weather conditions, the building's physical properties, occupancy patterns, and the comfort preferences of inhabitants is essential.
The ability to create such accurate and reliable flexibility profiles is being driven by state-of-the-art modeling techniques[42] that consider both physics and artificial intelligence techniques[45]. One prominent approach is the Grey-Box model, which provides a powerful hybrid approach between purely physical and data-driven methods. These

models use a simplified physical framework to represent a building's thermal properties, which is then continuously calibrated with real-world sensor data. On the other hand, Deep Reinforcement Learning (DRL)[43] offers another AI based approach that learns from the control strategy for an HVAC or DHW system through a process of digital trial and error. The latest trend is the Physics-Informed Neural Network (PINN)[44]. This new breed of model incorporates the fundamentals of thermodynamics into the neural network learning, allowing this new model to always be physically plausible in its predictions. The benefit of PINNs is their ability to create strong and accurate models, even with very sparse data, which is a great advancement in modeling thermal dynamics of buildings and water systems. Accurate flexibility profiles are essential to create highly reliable virtual power plants. Towards this direction the extraction of accurate demand side flexibility profiles will drive the development of the VPP application as well as the flexibility marketplace application in DIGITISE.

6.8.2 Technological Implementation and Deployment

Within DIGITISE, and as part of the modeling activities two distinct models are considered to address HVAC and DHW flexibility profiling.

Starting with HVAC flexibility modeling, a Gradient Boosting Machine (GBM) was developed to forecast baseline energy consumption and simulate demand response potential. The model was developed using high-resolution time-series datasets, which were acquired, cleaned, and synchronized. The data sample at 15-minute intervals:

- Building/System Data: HVAC power (kW), air temperature (°C), and thermostat setpoints (°C).
- External Context Data: Outdoor temperature (°C) and relative humidity (%).

To enhance the model's predictive power, a comprehensive feature engineering process was undertaken. Raw data was transformed into meaningful features designed to capture the system's operational dynamics:

- Temporal Features: Timestamps were converted into cyclical features, including hour of the day and day of the week, to model daily and weekly operational patterns.
- Thermal Dynamics Features: The temperature difference (ΔT) between outdoor and indoor environments was engineered as a primary driver of thermal load. Furthermore, lagged variables for power consumption and indoor temperature were created to provide the model with an understanding of the building's thermal inertia.
- Occupancy Related information: In the absence of direct occupancy data, robust occupancy proxies were developed by combining temporal features to distinguish between active and inactive

Then, a Gradient Boosting Machine (GBM), specifically the LightGBM implementation, was selected for its high performance and efficiency in handling tabular data. The implementation was executed in two primary stages: (i) A baseline model was successfully trained to predict the "business-as-usual" HVAC energy consumption. Using the engineered features as inputs and historical power consumption as the target variable, the model learned the dynamics of the problem. (ii) With a validated baseline model in place, it was then utilized as a reference to simulate and quantify flexibility. For any given period, the available power reduction ("shed" potential) was calculated, captured as the flexible load forecast. Once the GBM baseline model was validated in the configuration layer, it was serialized and transitioned to the execution layer utilized as a reference to simulate and quantify flexibility. Configured to run every 15 min, the engine runs the model to calculate the available power and power reduction ("shed" potential) for the next 2-3 hours in 15 min granularity.

For Domestic Hot Water (DHW) flexibility modeling, a Long Short-Term Memory (LSTM) network, is developed to forecast baseline energy consumption and simulate demand response potential similar to the case of HVAC. The model is developed using high-resolution time-series data, sampled at 15-minute intervals, which include:

- System Data: DHW heater power consumption (kW), internal water tank temperature (°C),
- Contextual Usage Data: Hot water draws events (liters or event markers) and the ambient temperature around the tank.

The functionality of the LSTM model is divided into two primary stages. First, a baseline model predicts the "business-as-usual" Domestic Hot Water (DHW) energy consumption. Using sequences of engineered features as inputs and historical power consumption as the target variable, the network captures the complex temporal dynamics of the system. Second, with the baseline established, this model serves as a reference to quantify flexibility. The available power reduction ("shed" potential) is calculated by assuming a "turn off" control action, making the available flexibility equal to the baseline power forecast. In its operational deployment, the validated and serialized LSTM model runs within the execution layer. The engine (similar to the model established for HVAC) is configured to execute every 15 minutes, using the model to calculate both the baseline power consumption and the available "shed" potential for the next 2-3 hours at a 15-minute granularity.

The performance of both models for the baseline forecast is to be evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The model's objective is to accurately forecast near-term power demand based on historical data and relevant features. The evaluation was conducted by comparing the model's predictions against a hold-out test dataset of actual consumption values.

For the current version and evaluation there are two core datasets considered for evaluation.

- For the HVAC system operation, the PLEIAData [46] is a comprehensive dataset designed to support a wide range of applications for smart buildings, with a focus on energy efficiency and user comfort.
- For the DHW system operation, the "Synthetic domestic hot water profile" dataset is considered with minute-by-minute profiles of domestic hot water (DHW) [47] usage for residential homes.

Early results are available only for the HVAC modeling as stated above, as the evaluation of the DHW model on the basis of the DHW dataset is still under evaluation. The target variable is the HVAC system's power demand, measured in kilowatts (kW), as the flexibility is then extracted from the baseline calculation.
The MAPE value is 0.1579 while the RMS=0.1781 kW for the specific dataset. The MAPE of 15.79% provides the relative error that is affected by the exact values of the dataset considered for training and testing/evaluation. The model is deemed successful and suitable for integration into the operational environment of the project though we have to further evaluate in DIGITISE data availability and cleanliness. The accuracy of the flexibility forecast will be validated during the demo period, where the model's predictions will be tested against the results of real-world control actions.

Funded by the European Union

7 DIGITISE Integrated framework

In this section, the scope is twofold (i) to present the details of the DIGITISE Data Space Management Layer, which functions as the orchestration and coordination component of the data space environment, and (ii) to present the integration details of the DIGITISE end to end framework.

7.1 DIGITISE Data Space integration and operation

As stated above, the focus is on the deployment of the different data space services as well as the execution of the different services stated above. As stated in D2.2, this layer ensures the seamless integration and operation of services and features provided across the other layers of the data spaces. It manages workflows, service execution, and resource coordination, serving as the control hub that brings coherence and efficiency to the overall data space operations. The details of the different components defined to support the data space operation are presented below.

7.1.1 Execution Master

7.1.1.1 Overview

The Data Execution Master is the singular orchestrator of all data processing pipelines within the DIGITISE ecosystem. It is responsible for running workflows on-demand to optimally transport data to its final, analysis-ready state. The Data Execution Master does not manipulate data but proactively orchestrates an intelligent sequence of a collection module, harmonization module, curation module. The Data Execution Master exercises substantial error recovery and error-handling options to execute and govern any data set in DIGITISE in a timely fashion. The Data Execution Master ultimately ensures the reliability, security, and governance of all data processing configurations in DIGITISE occur as intended.

7.1.1.2 Delivered Functionality

This section describes the delivered functionalities of the Data Execution Master as implemented in the first release. This was based upon the requirements defined in D2.1 and detailed functional design in D2.2. The delivery of these capabilities completes an end-to-end orchestration capability allowing for the automated execution of the entire data ingestion and preparation pipeline. The complete list of delivered core execution capabilities is:

- Execution of data collection (EM-01): The component orchestrates the gathering of raw data from various sources ensuring the initial step of the workflow is complete and accurate.

- Execution of data harmonization (EM-02): It triggers the harmonization process, which standardizes data formats and structures to enable seamless integration across diverse datasets.
- Execution of data curation (EM-03): It triggers the data curation, which clean, validate, and enrich data to enhance its overall quality, consistency, and usability.
- Execution of data anonymization (EM-04): It triggers the execution of the anonymization, which removes sensitive information to protect privacy while maintaining data integrity.

Following this, the data storage as well as management of metadata artefacts in real time is also supported by the execution engine. It is evident that the execution of each data collection process is in line with the different data governance services as mentioned in sections 3 and 4 to ensure the prompt handling of data in motion.

7.1.1.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- The first release supports the orchestration of a linear data processing pipeline and focuses on a predefined sequence of steps. Therefore, the limitation on the execution is on the limitations defined at the different steps as named above in section 3 and 4 (partial support of data collection methods, partial functionality on data curation/anonymization etc..).
- The reliability of the Data Execution Master is fundamentally dependent on the stability and performance of the individual components it orchestrates. Any changes to the functionality of these downstream services require corresponding updates to the orchestration logic to ensure seamless pipeline execution.
- Last but not least, error handling and recovery mechanisms during the execution have not been implemented yet in the 1st version and therefore, an attempt will be made to ensure the incorporation of the relevant mechanisms as part of the execution environment of DIGITISE.

7.1.2 Operations Monitor

7.1.2.1 Overview

The Data Operations Monitor complements the functionality of the execution engine acting as the central observability platform for the DIGITISE ecosystem, providing critical visibility into the health and performance of all data executions. Its objective is to confirm the proper functioning of data collection and governance processes by capturing and logging key operational metrics from the Data Execution Master. The service merges this information to produce on-demand insights and updates on data quality, execution completeness and timeliness, which data providers can easily access from a user-

interface that allows them to quickly troubleshoot problems while ensuring a higher velocity of dependable data quality.

7.1.2.2 Delivered Functionality

This section outlines the delivered functionalities of the Data Operations Monitor as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides a complete monitoring lifecycle, from logging raw operational data to presenting actionable quality insights and ensuring contractual compliance. The list of core delivered features is:

- Monitoring and logging (OM-01): A comprehensive logging mechanism has been delivered to capture essential metrics and events from data workflows, including execution status, completeness, and timeliness.
- Data quality insights (OM-02): The service generates regular "heartbeat" updates for all data collection pipelines and datasets, ensuring stakeholders remain informed about the current operational state of their data assets.
- User interface Viewer (OM-03): A user interface has been provided, allowing data providers to easily view metrics related to data health, execution details, and quality assessments to inform their data management strategies.

The component's comprehensive monitoring framework, as realized in this release, establishes a solid base for observability of the different data execution pipelines.

7.1.2.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- Apart from the monitoring system, a base user interface for viewing status is provided but new visual elements will be added in order to provide fine grained access to the logging information following execution.
- The accuracy and depth of the insights provided by the Data Operations Monitor are directly dependent on the richness of the logs and metrics generated by the upstream components, particularly the Data Execution Master. Maintaining consistent logging across all services is an ongoing operational requirement.
- While the first release provides robust monitoring, the alerting mechanisms has not been released. Future enhancements will focus on developing a configurable alerting system.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for users. In the next release, the aim is to evolve the component providing visual insights and more precise info about the execution of the different processes.

7.1.3 Resource Management

7.1.3.1 Overview

In DIGITISE's data space, the Resource Management Layer is their DevOps layer and is in charge of managing the resources of the supporting infrastructure. In this role, the primary role of the Resource Management Layer is to provide the stability, efficiency, and performance of the data space through managing calculations, storage, and network resources. This layer provides advanced resource scheduling to optimize workloads, constantly monitors the health and state of all layers for proactive management, and provides elasticity that uses the cloud infrastructure for dynamic scaling to accommodate workloads based on real-time demand. The Resource Management Layer provides ongoing oversight and coordination functions that ensures a responsive environment for the data operations that occur in the data space environment.

7.1.3.2 Delivered Functionality

This section outlines the delivered functionalities of the Resource Management Layer as implemented in the first release, based on the requirements defined in D2.1 and the detailed functional design in D2.2. The implementation of these features provides a complete lifecycle for resource management, from initial orchestration and allocation through to real-time monitoring and dynamic scaling. The list of core delivered features is:

- Orchestration of data ecosystem resources (RM-01): The component manages the interaction and interfacing of the critical resources (compute, storage, network) for separate processes to execute in the data ecosystem.
- Resources allocation (RM-02): an allocation mechanism which is provided to efficiently leverage storage and compute resources through the distribution of current workloads.
- Real-time monitoring (RM-03): The layer continuously tracks the status and health of resources, with alert systems to notify administrators of failures or performance issues, allowing for swift intervention.

7.1.3.3 Considerations, Assumptions, and Constraints

In the initial release, certain constraints have been established that may affect the capabilities of the DIGITISE Data Space.

- Moreover, the logging mechanism needs to be enhanced in order to have full monitoring at DevOps level of the different processes
- Last but not least the resource management function needs to be updated taking into account the updates/enhancements to be performed to the different services as specified in Section 3 and 4, reflecting also the execution through the Execution master as specified above.

These constraints define the scope of functionalities available in the current version of DIGITISE and are important considerations for the next release as the aim is to evolve the component from a static to a more dynamic resource management system.

7.1.4 Technology Stack and development details

Following the detailed presentation of the different components that consist of the Data Operation Layer of DIGITISE, an overview of the different technologies and frameworks considered for the delivery of the services is presented below.

Library	Version	License
NodeJS	18	MIT
Vue.js	2.7	MIT
TailwindCSS	2	MIT
Kubernetes	-	Apache License 2.0
KEDA	-	Apache License 2.0
Prometheus	-	MIT
Kafka	-	Apache License 2.0
PostgreSQL	-	PostgreSQL License (similar to BSD/MIT)
Redis	-	3-clause BSD License
MinIO	-	Apache License 2.0
MongoDB	-	Apache License 2.0
Vault	-	Mozilla Public License 2.0

TABLE 8 DIGITISE DATA OPERATION TECHNOLOGIES

This section provides a high-level overview of the key technologies and frameworks used for the implementation of the Data Operation Layer.

- The Resource Management Layer is built on Kubernetes as the underlying container orchestration platform for deploying, managing, and scaling all system components. To further enable more advanced autoscaling for event-driven workloads, KEDA is integrated with Kubernetes to allow Kubernetes resources to be used or removed based on workloads in real-time (e.g. message queue length, etc). Secure management of all accesses within the cluster are handled by Vault, guarding secrets, certificates, and credentials.
- The Execution Master service functions as the workflow engine, orchestrating data pipelines execution. Its backend is developed in NodeJS. It leverages a messaging system based on Kafka to manage state and trigger sequential tasks in a resilient, asynchronous manner. The definitions and current state of active workflows are persisted in PostgreSQL. The data as extracted from execution process are stored in MinIO and MongoDB.

- The Data Operations Monitor stack provides comprehensive monitoring of the entire ecosystem. Prometheus is used as the central metrics collection engine, scraping performance and health data from all services. Monitoring interfaces are developed using a frontend stack of Vue.js /TailwindCSS to provide tailored insights for data providers. The backend of the service is developed in NodeJS.

7.2 DIGITISE integration

For a smooth integration of the WP4 and WP5 services and tools into the DIGITISE ecosystem, a set of monitoring components will be employed in order to collect, analyze and visualize relevant system performance information from the different premises where the DIGITISE applications are deployed. This will address the need for ensuring reliability, scalability, and efficient resource utilization, across distributed application environments.

For this purpose, the solution to be followed in the DIGITISE project includes the setup of Node Exporters on the different servers to be monitored and a centralized Prometheus deployment accompanied by a Grafana visualization dashboard.

With the use of Node Exporter at each server, where an application will be deployed, several system metrics, like CPU utilization, memory usage, disk I/O statistics, network traffic and more, will be gathered and formatted in a way that can be easily read and stored by Prometheus.

A Prometheus application will be centrally installed and will be configured so that it pulls the aforementioned measured metrics at regular intervals from the different clients where the exporters are installed and stores it in its dedicated database for real-time analysis or later use with a number of inspection tools.

The DIGITISE integration and monitoring mechanism is completed by the exploitation of Grafana, a tool that provides intuitive and customizable visualizations from various data sources. In DIGITISE, Grafana will enable its users to gain insights into the performance and health of the different applications and tools integrated in the DIGITISE environment through dynamic dashboards from a centralized access point.

Since the DIGITISE integration task is heavily based on the deployment of the different WP4 and WP5 tools and applications, it is planned to begin after the completion of the 1st version of the applications. The integration rollout will include the following phases:

- The first step towards the realization of the full set of monitoring mechanisms is to have the different implementations deployed and functional, after the 1st version

release. Based on the decided deployment method of each application (Kubernetes, Docker, native installations, etc.) the Node Exporter integration technique will be defined to collect the desired metrics on machine (server) level and/or i.e. on container level (in the case of the dockerization solution, where a cAdvisor component will also need to be installed alongside Node Exporter).

- After the definition of metrics' collection method, the different values to be monitored will be decided along with relevant partners, considering their needs for better understanding the performance and health status of their applications that would effectively lead them to proactive actions, and errors or bottlenecks prevention and management. An initial set of metrics will be defined, which will be refined if needed after the respective findings.

- The centrally installed Prometheus application will be set up and gradually configured, until the whole set of WP4 and WP5 tools are equipped with the Node Exporter and some initial, indicative metrics are successfully scraped and stored in the Prometheus database. When the integration process will have reached a mature phase, an alerting mechanism could also be set up, in case any anomalies are observed, initially for notifying UBITECH as the main integration partner, and at a later stage, considering the customization capabilities of the tool, to explore more personalized options per partner/application deployment.

- After the successful configuration and testing of the Prometheus application, where collected values will be available for analysis, Grafana will be integrated into the system, and several informative graphs will be defined based on the gathered metrics and the operations that partners wish to observe. One or more intuitive dashboards will be built, and if needed will be shared with the respective partners.

The different technologies to be considered for the integration components described above are also listed in the table below:

Library	Version	License
Prometheus	3.5.0	Apache License 2.0
Node exporter	1.9.1	Apache License 2.0
Grafana	12.1	AGPL-3.0 license

TABLE 9 DIGITISE DATA INTEGRATION MONITORING TECHNOLOGIES

8 Summary and Conclusions

This report outlines the progress made within Work Package 3, "DIGITISE Reference Energy Data Space Implementation," leading up to the deployment of the draft version of the DIGITISE Data Space at month 15 of the DIGITISE project.

The process of developing the DIGITISE data model occurred during Task 3.1 of the project, and was done in a structured manner. The first stage of the project was a state-of-the-art (SOTA) investigation, which provided a comprehensive overview of existing standards and best practice. For the subsequent stage, a detailed landscape review was completed in order to assess relevant data availability and public data sources, as well as stakeholder requirements, and interoperability challenges in the energy industry. Following this, the data model was developed, which effectively grouped the insights and learning from the previous stages to develop a model that is comprehensive, flexible, addresses the contemporary and future needs, and scalable. This thorough and consistent methodical analysis assured that the data model is understandable, integrates seamlessly, and harmonises via the interaction of multiple data sources and applications in the DIGITISE ecosystem.

In accordance with the D2.2 document, the progress for each building block, status and implementation methods up until M15, technology stack, relevant assumptions and limitations have been documented sequentially and individually. Additionally, the report captures the deployment information for the first release of the DIGITISE Data Space, and its support documents as necessary. In addition to its core data management functionality, DIGITISE provides a full suite of analytics capabilities that generate insight and facilitate higher-level operational decision-making. Within DIGITISE the analytics extensions use artificial intelligence and machine learning to analyze, track and preform trend analysis of data from batch, real-time and streaming workloads, maximizing value from the data assets available to them. The analytics environment within DIGITISE will provide a scalable and configurable environment to respond to increased data quantities and analytical needs for the series of pre-trained energy and non-energy-based models provided in the project.

The document further elaborates on the integration strategy, which is closely aligned with agile development methodologies throughout the software delivery cycles. The analysis addresses the integration needs both at the level of the data space but also beyond considering the integration of the DTs and the energy apps of the project. While there is significant progress towards the delivery of the data space environment, all DIGITISE digital solutions and the energy services marketplace (delivered in WP4 and WP5) will be integrated, introducing the energy added-value services dimension; integration details for the rest of the applications are to be performed in the next period (thus only the methodological process and the steps are outlined in this version of the document).

Finally, Annex I of this report documents the DIGITISE Network of Data Models (Energy, Health, Finance), developed as part of Task 3.1. The DIGITISE Network of Data Models is integrated into the initial release of the DIGITISE Data Space and will be maintained and expanded using the available CIM Network management functionalities.

Towards the final release of the DIGITISE Data Platform at month 30 features partially implemented in the draft release will be enhanced, and those omitted from the first release will be fully implemented in accordance with the project architecture outlined in deliverable D2.2. Furthermore, feedback from technical partners integrating their solutions in WP4/5, along with input from demo partners conducting demonstration activities in WP6, "Demonstration and Impact Assessment," will be utilized to refine features and enhance the overall user experience.

9 References

[1]. Otto, B., Auer, S., et al. (2022). *IDSA Reference Architecture Model 4.0*. International Data Spaces Association.

[2]. Platschek, H., et al. (2021). *Gaia-X Architecture Document*. Gaia-X, a European project for data.

[3]. ETSI. (2017). *ETSI TR 103 264 - SAREF (Smart Applications REFerence ontology)*. European Telecommunications Standards Institute.

[4]. W3C. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. W3C Recommendation.

[5]. SYNERGY Project Consortium. (2021). *D3.3 – Big Data Analytics Platform Final Release*. H2020-SYNERGY Project Deliverable.

[6]. Modbus Organization. (2012). *Modbus Application Protocol Specification V1.1b3*.

[7]. Open Charge Alliance. (2020). *Open Charge Point Protocol (OCPP) 2.0.1*.

[8]. Hu, V. C., Ferraiolo, D., Kuhn, R., et al. (2014). *Guide to Attribute-Based Access Control (ABAC) Definition and Considerations*. NIST Special Publication 800-162.

[9]. International Data Spaces Association. (2022). *IDSA Position Paper: An Interoperability Framework for the Energy Data Space*. IDSA.

[10]. CEN-CENELEC-ETSI Smart Grid Coordination Group. (2012). *Smart Grid Reference Architecture*.

[11]. International Electrotechnical Commission. (2021). *IEC 61970/61968/62325 Series - Common Information Model (CIM)*.

[12]. International Electrotechnical Commission. (2020). *IEC 61850 Series - Communication networks and systems for power utility automation*.

[13]. International Electrotechnical Commission. (2021). *IEC 61850-7-420:2021 - Communication networks and systems for power utility automation – Part 7-420: Basic communication structure – Distributed energy resources and distribution automation logical nodes*.

[14]. International Electrotechnical Commission. (2018). *IEC TR 61850-90-8:2018 - Communication networks and systems for power utility automation – Part 90-8: Object models for electric vehicle charging*.

[15]. Institute of Electrical and Electronics Engineers. (2018). *IEEE 2030.5-2018 - IEEE Standard for Smart Energy Profile Application Protocol*.

[16]. International Electrotechnical Commission. (2024). *IEC 62746-4:2024 - Systems interface between customer energy management system and the power management system – Part 4: Demand Side Resource Interface*.

[17]. OpenADR Alliance. (2019). *OpenADR 2.0b Profile Specification*.

[18]. Open Charge Alliance. (2020). *Open Charge Point Protocol (OCPP) 2.0.1*. (Standardized as IEC 63110).

[19]. European Telecommunications Standards Institute. (2021). *ETSI TS 103 264 - SmartM2M; Smart Applications; Reference Ontology (SAREF) Family*.
[20]. Balaji, B., et al. (2018). "Brick: A Uniform Metadata Schema for Buildings". In *Proceedings of the 5th ACM International Conference on Systems for Built Environments (BuildSys '18)*.
[21]. Project Haystack Organization. (n.d.). *Project Haystack Documentation*. Retrieved from https://project-haystack.org/
[22]. Rasmussen, K., et al. (2021). *Building Topology Ontology (BOT)*. W3C Community Group Final Report.
[23]. International Organization for Standardization. (2018). *ISO 19650 Series - Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM)*.
[24]. International Organization for Standardization. (2024). *ISO 16739-1:2024 - Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries — Part 1: Data schema*.
[25]. Connectivity Standards Alliance. (2022). *Matter Specification*.
[26]. American Society of Heating, Refrigerating and Air-Conditioning Engineers. (2022). *ANSI/ASHRAE Standard 135-2022 - BACnet—A Data Communication Protocol for Building Automation and Control Networks*.
[27]. OPC Foundation. (2017). *OPC Unified Architecture Specification*.
[28]. International Energy Agency, Energy in Buildings and Communities Programme (IEA EBC). (2023). *Annex 79: Occupant-Centric Building Design and Operation - Final Report*.
[29]. International Energy Agency, Energy in Buildings and Communities Programme (IEA EBC). (2017). *Annex 66: Definition and Simulation of Occupant Behavior in Buildings - Final Report*.
[30]. Chen, Y., & Hong, T. (2018). "obXML: A schema for exchanging occupant behavior models in building simulation". *Journal of Building Performance Simulation*, 11(1), 49-63.
[31]. Donkers, A., & de Rijke, M. (2021). "The Occupant Feedback Ontology (OFO): A standardised vocabulary for occupant feedback". In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '21)*.
[32]. Digital Construction Ontologies. (n.d.). *Indoor Air Quality Ontology*. Retrieved from https://digitalconstruction.github.io/IndoorAirQuality/
[33]. Alhamwi, A., et al. (2024). "IAQ-BIM Ontology: A Semantic Web-Based Approach for Integrating Indoor Air Quality, Building Energy Performance, and Health Data". *Sustainability*, 16(13), 5677.
[34]. International Organization for Standardization. (2004-2022). *ISO 16000 Series - Indoor air*.

[35]. American Society of Heating, Refrigerating and Air-Conditioning Engineers. (2022). *ANSI/ASHRAE Standard 62.1 and 62.2 - Ventilation and Acceptable Indoor Air Quality*.

[36]. European Committee for Standardization (CEN). (2019). *EN 16798-1:2019 - Energy performance of buildings - Part 1: Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics*.

[37]. buildingSMART alliance. (2012). *Construction-Operations Building information exchange (COBie) - National BIM Standard-United States Version 2*.

[38]. Enterprise Data Management Council (EDMC). (n.d.). *Financial Industry Business Ontology (FIBO)*. Retrieved from https://spec.edmcouncil.org/fibo/

[39]. XBRL International, Inc. (n.d.). *The XBRL Standard*. Retrieved from https://www.xbrl.org/

[40]. HOMER Energy by UL Solutions. (n.d.). *HOMER Software for distributed generation and microgrid design*. Retrieved from https://www.homerenergy.com/

[41]. National Renewable Energy Laboratory (NREL). (n.d.). *REopt: Renewable Energy Integration & Optimization Platform*. Retrieved from https://reopt.nrel.gov/

[42]. Cui, B., Fan, C., Mofid, S. A., & Augenbroe, G. (2021). "A review of data-driven building thermal models: Recent developments, strengths and weaknesses". Renewable and Sustainable Energy Reviews, 145, 111059.

[43]. Wang, Z., & Hong, T. (2020). "Reinforcement learning for building controls: A review". Automation in Construction, 115, 103180.

[44]. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". Journal of Computational Physics, 378, 686-707.

[45]. Drgoňa, J., et al. (2023). "Data-driven building modeling and control: A review of the state-of-the-art, current challenges, and future outlooks". Annual Reviews in Control, 56, 100891.

[46]. Ibarra, A.M., González-Vidal, A. & Skarmeta, A. PLEIAData: consumption, HVAC, temperature, weather and motion sensor data for smart buildings applications. Sci Data 10, 118 (2023). https://doi.org/10.1038/s41597-023-02023-3

[47]. https://scholardata.sun.ac.za/articles/software/Synthetic_domestic_hot_water_profile_generator/12173886

[48]. Zheng, M. Enhancing energy efficiency in HVAC systems through precise heating load forecasting and advanced optimization algorithms. (2024)

[49]. Aghbalou, N., Charki, A., Errousso, H., Filali, Y. Ensemble Learning Method for Forecasting HVAC System Demand. (2024)

[50]. Salem KM, Rey-Hernández JM, Elgharib AO, Rey-Martínez FJ. Optimizing Energy Forecasting Using ANN and RF Models for HVAC and Heating Predictions. Applied Sciences. 2025

[51]. Sirine Maalej, Zoubeir Lafhaj, Jean Yim, Pascal Yim, Colin Noort. Prediction of HVAC System Parameters Using Deep Learning. 12th eSim Building Simulation Conference. 2022

[52]. Xiao Z, Yu L, Zhang H, Zhang X, Su Y. HVAC Load Forecasting Based on the CEEMDAN-Conv1D-BiLSTM-AM Model. Mathematics. 2023

[53]. A. Pérez-Fargallo, D. Bienvenido-Huertas, S. Contreras-Espinoza, L. Marín-Restrepo, Domestic hot water consumption prediction models suited for dwellings in central-southern parts of Chile, Journal of Building Engineering, 2022

[54]. Gelažanskas L, Gamage KAA. Forecasting Hot Water Consumption in Residential Houses. Energies. 2015

[55]. Bayle R, Reyboz M, Lomet A, Cook V, Mermillod M. Continuously Learning Prediction Models for Smart Domestic Hot Water Management. Energies. 2024

[56]. Louis-Gabriel Maltais, Louis Gosselin, Predictability analysis of domestic hot water consumption with neural networks: From single units to large residential buildings, Energy, 2021

[57]. Ibrahim Ali Kachalla, Christian Ghiaus. Forecasting Hot Water Consumption Demand for Residential Dwellings Using Hybrid Regression Technique. 15th IEEE International Conference on Green Energy and Smart Systems, 2024

[58]. Fan, Longmao, Jianing Li, and Xiao-Ping Zhang. "Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition." CSEE Journal of Power and Energy Systems 6.3 (2020): 563-571.

[59]. Michalakopoulos, Vasilis, Elissaios Sarmas, Ioannis Papias, Panagiotis Skaloumpakas, Vangelis Marinakis, and Haris Doukas. "A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs." Applied Energy 361 (2024): 122943.

[60]. Kong, Weicong, Zhao Yang Dong, Youwei Jia, David J. Hill, Yan Xu, and Yuan Zhang. "Short-term residential load forecasting based on LSTM recurrent neural network." IEEE transactions on smart grid 10, no. 1 (2017): 841-851.

[61]. Stephen, Bruce, Xiaoqing Tang, Poppy R. Harvey, Stuart Galloway, and Kyle I. Jennett. "Incorporating practice theory in sub-profile models for short term aggregated residential load forecasting." IEEE Transactions on Smart Grid 8, no. 4 (2015): 1591-1598.

[62]. Yilmaz, Selin, Jonathan Chambers, and Martin Kumar Patel. "Comparison of clustering approaches for domestic electricity load profile characterisation-Implications for demand side management." Energy 180 (2019): 665-677.

[63]. Okereke, George Emeka, Mohamed Chaker Bali, Chisom Nneoma Okwueze, Emmanuel Chukwudi Ukekwe, Stephenson Chukwukanedu Echezona, and Celestine Ikechukwu Ugwu. "K-means clustering of electricity consumers using time-domain features from smart meter data." Journal of Electrical Systems and Information Technology 10, no. 1 (2023): 2.

[64]. Zhang, Liping, Song Deng, and Shiyue Li. "Analysis of power consumer behavior based on the complementation of K-means and DBSCAN." In 2017 IEEE conference on energy internet and energy system integration (EI2), pp. 1-5. IEEE, 2017.

[65]. Fu, Jiasha, Shan Hu, Xin He, Shunsuke Managi, and Da Yan. "Identifying residential building occupancy profiles with demographic characteristics: using a national time use survey data." Energy and Buildings 277 (2022): 112560.

[66]. Wilke, Urs, Frédéric Haldi, Jean-Louis Scartezzini, and Darren Robinson. "A bottom-up stochastic model to predict building occupants' time-dependent activities." Building and Environment 60 (2013): 254-264.

[67]. Buttitta, Giuseppina, William Turner, and Donal Finn. "Clustering of household occupancy profiles for archetype building models." Energy Procedia 111 (2017): 161-170.

[68]. Vassiljeva, Kristina, Margarita Matson, Andrea Ferrantelli, Eduard Petlenkov, Martin Thalfeldt, and Juri Belikov. "Data-driven occupancy profile identification and application to the ventilation schedule in a school building." Energies 17, no. 13 (2024): 3080.

[69]. Zhan, Sicheng, and Adrian Chong. "Building occupancy and energy consumption: Case studies across building types." Energy and Built Environment 2, no. 2 (2021): 167-174.

[70]. Song, Ying, Fubing Mao, and Qing Liu. "Human comfort in indoor environment: a review on assessment criteria, data collection and data analysis methods." Ieee Access 7 (2019): 119774-119786.

[71]. Wu, Siyu, and Jian-Qiao Sun. "A physics-based linear parametric model of room temperature in office buildings." Building and Environment 50 (2012): 1-9.

[72]. Kim, Joyce, Yuxun Zhou, Stefano Schiavon, Paul Raftery, and Gail Brager. "Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning." Building and environment 129 (2018): 96-106.

[73]. Boutahri, Youssef, and Amine Tilioua. "Machine learning-based predictive model for thermal comfort and energy optimization in smart buildings." *Results in Engineering* 22 (2024): 102148.

[74]. Wei, Wenjuan, Olivier Ramalho, Laeticia Malingre, Sutharsini Sivanantham, John C. Little, and Corinne Mandin. "Machine learning and statistical models for predicting indoor air quality." Indoor Air 29, no. 5 (2019): 704-726.

[75]. Wei, Shuangyu, Paige Wenbin Tien, Tin Wai Chow, Yupeng Wu, and John Kaiser Calautit. "Deep learning and computer vision based occupancy CO2 level prediction for demand-controlled ventilation (DCV)." Journal of Building Engineering 56 (2022): 104715.

[76]. Drewil, Ghufran Isam, and Riyadh Jabbar Al-Bahadili. "Air pollution prediction using LSTM deep learning and metaheuristics algorithms." Measurement: Sensors 24 (2022): 100546.
[77]. Elmaz, Furkan, Reinout Eyckerman, Wim Casteels, Steven Latré, and Peter Hellinckx. "CNN-LSTM architecture for predictive indoor temperature modeling." Building and Environment 206 (2021): 108327.

10 Annex I

In this section we provide more details about the data model defined in the project. As stated above, the model is following an ER approach where the different concepts characterized by attributes and relations among the different concepts. For IP reasons no detailed list of relationships and attributes are defined, but we report the total number per concept in the following table.

ID	Concept	Category	Related Concepts	Related Attributes
1	AirConditioningSystem	Buildings & IoT	11	26
2	AirConditioningSystemControlAction	Buildings & IoT	7	21
3	Boiler	Buildings & IoT	12	23
4	BoilerControlAction	Buildings & IoT	10	8
5	Building	Buildings & IoT	19	37
6	BuildingMeasurements	Buildings & IoT	1	9
7	BuildingSpace	Buildings & IoT	21	14
8	BuildingStorey	Buildings & IoT	7	11
9	BuildingZone	Buildings & IoT	17	8
10	ChillerDevice	Buildings & IoT	11	13
11	ChillerDeviceControlAction	Buildings & IoT	6	8
12	DomesticHotWaterSystem	Buildings & IoT	12	10
13	DomesticHotWaterSystemControlAction	Buildings & IoT	6	8
14	ElectricAppliance	Buildings & IoT	13	11
15	ElectricApplianceControlAction	Buildings & IoT	8	7
16	Gateway	Buildings & IoT	4	7
17	HeatPump	Buildings & IoT	12	14
18	HeatPumpControlAction	Buildings & IoT	6	8
19	LightingDevice	Buildings & IoT	10	16
20	LightingDeviceControlAction	Buildings & IoT	7	10
21	Outlet	Buildings & IoT	11	13
22	SmartAppliance	Buildings & IoT	8	9
23	SmartApplianceControlAction	Buildings & IoT	6	7
24	SpaceHeatingDevice	Buildings & IoT	10	17
25	SpaceHeatingDeviceControlAction	Buildings & IoT	6	8
26	Address	Cross-Cutting Roles & Systems	2	16
27	Aggregator	Cross-Cutting Roles & Systems	11	6
28	AggregatorPortfolio	Cross-Cutting Roles & Systems	19	10
29	BalancingResponsibleParty	Cross-Cutting Roles & Systems	8	6
30	BalancingServiceProvider	Cross-Cutting Roles & Systems	8	6
31	DemandResponseEvent	Cross-Cutting Roles & Systems	14	12
32	DemandResponseEventSignal	Cross-Cutting Roles & Systems	16	15

33	DemandResponseReport	Cross-Cutting Roles & Systems	17	11
34	DemandResponseReportReading	Cross-Cutting Roles & Systems	17	9
35	Device	Cross-Cutting Roles & Systems	11	16
36	DeviceControlEvent	Cross-Cutting Roles & Systems	17	5
37	DeviceControlEventAction	Cross-Cutting Roles & Systems	2	11
38	DeviceControlStatus	Cross-Cutting Roles & Systems	17	7
39	EnergyDemandMeasurements	Cross-Cutting Roles & Systems	20	35
40	EnergyMarket	Cross-Cutting Roles & Systems	14	20
41	EnergyMarketOperator	Cross-Cutting Roles & Systems	3	6
42	EnergyServiceCompany	Cross-Cutting Roles & Systems	3	6
43	Event	Cross-Cutting Roles & Systems	2	9
44	FacilityManager	Cross-Cutting Roles & Systems	4	6
45	Flexibility	Cross-Cutting Roles & Systems	20	11
46	FlexibilityMarket	Cross-Cutting Roles & Systems	14	20
47	FlexibilityMarketOperator	Cross-Cutting Roles & Systems	3	6
48	Incident	Cross-Cutting Roles & Systems	8	11
49	IncidentLog	Cross-Cutting Roles & Systems	8	5
50	KeyPerformanceIndicator	Cross-Cutting Roles & Systems	19	5
51	KeyPerformanceIndicatorValue	Cross-Cutting Roles & Systems	2	8
52	LoadResponse	Cross-Cutting Roles & Systems	5	9
53	LocalEnergyCommunity	Cross-Cutting Roles & Systems	15	7
54	LocalEnergyCommunityPortfolio	Cross-Cutting Roles & Systems	5	1
55	Location	Cross-Cutting Roles & Systems	2	25
56	Measurement	Cross-Cutting Roles & Systems	16	7
57	MeteringSystem	Cross-Cutting Roles & Systems	7	8
58	NetworkSwitch	Cross-Cutting Roles & Systems	5	10
59	Offer	Cross-Cutting Roles & Systems	8	5
60	OfferOption	Cross-Cutting Roles & Systems	3	7
61	Order	Cross-Cutting Roles & Systems	11	9
62	Period	Cross-Cutting Roles & Systems	4	25
63	Prosumer	Cross-Cutting Roles & Systems	17	11
64	Request	Cross-Cutting Roles & Systems	6	7

65	Retailer	Cross-Cutting Roles & Systems	8	6
66	RetailerPortfolio	Cross-Cutting Roles & Systems	18	6
67	Schedule	Cross-Cutting Roles & Systems	9	7
68	Settlement	Cross-Cutting Roles & Systems	7	11
69	Status	Cross-Cutting Roles & Systems	2	14
70	TroubleTicket	Cross-Cutting Roles & Systems	11	7
71	VirtualPowerPlant	Cross-Cutting Roles & Systems	6	6
72	WeatherMeasurement	Cross-Cutting Roles & Systems	2	25
73	WeatherStation	Cross-Cutting Roles & Systems	7	5
74	chargePointOperator	EV Charging & e-Mobility	13	6
75	chargePointOwner	EV Charging & e-Mobility	14	6
76	chargeSession	EV Charging & e-Mobility	8	5
77	ElectricVehicle	EV Charging & e-Mobility	14	14
78	EVChargingPlatform	EV Charging & e-Mobility	2	4
79	EVChargingStation	EV Charging & e-Mobility	12	18
80	EVChargingStationControlAction	EV Charging & e-Mobility	8	9
81	EVUser	EV Charging & e-Mobility	13	6
82	Battery	Generation/Storage Systems	9	24
83	BatteryControlAction	Generation/Storage Systems	9	8
84	BiomassPlant	Generation/Storage Systems	9	22
85	CombinedHeatingPowerSystem	Generation/Storage Systems	11	16
86	Electrolyzer	Generation/Storage Systems	9	23
87	EnergyGenerationMeasurements	Generation/Storage Systems	17	115
88	EnergyStorageMeasurements	Generation/Storage Systems	4	17
89	FuelCell	Generation/Storage Systems	9	23
90	Generator	Generation/Storage Systems	8	14
91	HydrogenTank	Generation/Storage Systems	9	12
92	HydroPowerSystem	Generation/Storage Systems	6	8
93	Inverter	Generation/Storage Systems	11	23
94	PhotovoltaicSystem	Generation/Storage Systems	11	32
95	PlantOperator	Generation/Storage Systems	7	6
96	PowerPlant	Generation/Storage Systems	7	12

97	RenewableEnergySource	Generation/Storage Systems	13	5
98	RenewableEnergySourceOperator	Generation/Storage Systems	10	6
99	SolarThermal	Generation/Storage Systems	13	17
100	SolarThermalControlAction	Generation/Storage Systems	8	8
101	WindTurbine	Generation/Storage Systems	11	23
102	ACLine	Grid	9	33
103	Branch	Grid	5	12
104	Bus	Grid	4	11
105	ConnectivityNode	Grid	8	7
106	DistributionSystemOperator	Grid	5	5
107	DistributionSystemOperatorPortfolio	Grid	5	8
108	Grid	Grid	11	12
109	Impedance	Grid	1	6
110	Load	Grid	2	12
111	Outage	Grid	13	25
112	OutageLog	Grid	12	7
113	PowerTransformer	Grid	5	42
114	SCADA	Grid	3	7
115	Substation	Grid	9	11

ID	Concept	Category	Related Concepts	Related Attributes
1	Sensor	Cross-Cutting Roles & Systems	14	10
2	SensingMeasurement	Cross-Cutting Roles & Systems	12	30
3	HealthProblem	Indoor Air Quality (IAQ) & Health	2	12
4	Humidifier	Indoor Air Quality (IAQ) & Health	5	8
5	HumidifierControlAction	Indoor Air Quality (IAQ) & Health	10	17
6	IndoorAirQuality (IAQ)	Indoor Air Quality (IAQ) & Health	4	9
7	IndoorEnvironmentHealth	Indoor Air Quality (IAQ) & Health	8	25
8	MedicalDevice	Indoor Air Quality (IAQ) & Health	17	16
9	MedicalDeviceControlAction	Indoor Air Quality (IAQ) & Health	10	9
10	PollutantConcentration	Indoor Air quality (IAQ) & Health	1	12
11	VentilationSystem	Indoor Air quality (IAQ) & Health	13	17
12	VentilationSystemControlAction	Indoor Air quality (IAQ) & Health	8	10
13	AvoidableDisposition	Occupant Comfort & Behavior	2	6
14	Occupancy	Occupant Comfort & Behavior	4	18
15	OccupancyActivity	Occupant Comfort & Behavior	1	6
16	OccupancyProfile	Occupant Comfort & Behavior	1	8
17	Occupant	Occupant Comfort & Behavior	2	15
18	OccupantBehavior	Occupant Comfort & Behavior	1	5
19	SoundComfort	Occupant Comfort & Behavior	6	15
20	SoundLevel	Occupant Comfort & Behavior	3	15

21	ThermalComfort	Occupant Comfort & Behavior	8	26
22	VisualComfort	Occupant Comfort & Behavior	6	12
23	AccessControl	Physical Building Security	4	8
24	AlarmIntegration	Physical Building Security	3	5
25	AttackMethod	Physical Building Security	2	7
26	BurglaryResistance	Physical Building Security	2	6
27	IntrusionDetection	Physical Building Security	2	5
28	Physical Security	Physical Building Security	2	5
29	ThreatLevel	Physical Building Security	4	6

ID	Concept	Category	relatedConcepts	relatedAttributes
1	Arrangement	Agreements & Contracts	4	8
2	Contract	Agreements & Contracts	5	7
3	Lease	Agreements & Contracts	2	6
4	Loan	Agreements & Contracts	2	7
5	Mortgage	Agreements & Contracts	2	8
6	RepaymentSchedule	Agreements & Contracts	2	9
7	FinancialIndicator	Financial Instruments & Valuation	4	10
8	FinancialInstrument	Financial Instruments & Valuation	3	8
9	FinancialProduct	Financial Instruments & Valuation	4	6
10	InterestRate	Financial Instruments & Valuation	2	7
11	ProjectValuation	Financial Instruments & Valuation	3	8
12	Security	Financial Instruments & Valuation	2	8
13	FinancialBody	Project Roles & Entities	5	9
14	Investor	Project Roles & Entities	3	9
15	Lender	Project Roles & Entities	2	8
16	Manufacturer	Project Roles & Entities	2	8
17	Project	Project Roles & Entities	2	15
18	ProjectDeveloper	Project Roles & Entities	2	10
19	FinancialRegulation	Regulatory	1	6
20	RegulatoryAgency	Regulatory	1	5
21	Incentive	Utility, Tariffs & Incentives	1	10
22	TariffProfile	Utility, Tariffs & Incentives	5	15
23	TariffRateComponent	Utility, Tariffs & Incentives	5	9
24	UtilityBill	Utility, Tariffs & Incentives	3	10
25	UtilityTariff	Utility, Tariffs & Incentives	3	8

More than 150 concepts are defined to set the basis for the model with more than 1000 attributes to be defined to characterize the different concepts. We need to point out that the definition of entities and attributes is a continuous process to be performed in the project in order to ensure that all data elements are well characterized in the DIGITISE project.

Note: while the energy and health data models are rather mature, further analysis on the Financial data model is required to be performed during the next period and results to be reported in the 2nd version of the deliverable.