BD4NRG
Big Data for Next Generation Energy

D4.2: Report on Existing Methods, Tools and Prototype Implementations to realize the Semantic Interoperability Toolbox, Framework and Platform

WP4 – Data Modelling & Open Modular AI-based edge-level Analytics

February 2022

www.bd4nrg.eu

Disclaimer

Copyright Message

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613

Grant Agreement Number: 872613 Acronym: BD4NRG

Full Title	Big Data for Next Generation Energy		
Topic	DT-ICT-11-2019 Big data solutions for energy		
Funding scheme	H2020- IA: Innovation Action		
Start Date	January 2021	Duration	36
Project URL	http//www.bd4nrg.eu		
Project Coordinator	ENG		
Deliverable	D4.2 – Report on Existing Methods Tools and Prototype Implementations to realize the Semantic Interoperability Toolbox, Framework and Platform		
Work Package	WP4 – System Requirements and Specifications		
Delivery Month (DoA)	M14	Version	1.0
Actual Delivery Date	01/12/2022		
Nature	Report	Dissemination Level	Public
Lead Beneficiary	TNO		
Authors	Harrie Bastiaansen (TNO), Michiel Stornebrink (TNO), Arjan Stoter (TNO), Wouter van den Berg (TNO), Anastasia Anagnostopoulou (NTUA), Efthimios Bothos (NTUA), Babis Magoutas (NTUA)		
Quality Reviewer(s):	Antonello Monti (RWTH) Marzia Mammina (ENG)		
Keywords	Semantic Interoperability, Broker, Meta-data Catalogue, Vocabulary Hub, Data Transformation, Data Validation.		

Preface

BD4NRG focuses on addressing emerging challenges in big data management for the energy sector with an innovative open holistic solution for smart grid-tailored, near real time, energy-specific and AI-based open Big Data Analytics modular framework. The vision is to deliver holistic services for techno-economic optimal management of Electric Power and Energy Systems value chain. Services range from optimal risk assessment for energy efficiency investments planning, to optimized management of grid and non-grid owned assets, improved efficiency, and reliability of electricity networks operation, while at the same time contributing to achieve fair energy prices to the consumers and laying the foundations for an EU-level energy-tailored data sharing economy. The BD4NRG Toolbox will be implemented and validated in real-life pilots in 12 large-scale demo sites across 8 countries for:

- Increasing the efficiency and reliability of the electricity network BD-4-NET
- Optimising the management of assets connected to the grid BD-4-DER

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- De-risking investments in energy efficiency and increasing the efficiency and comfort of buildings
  BD-4-ENEF

Who We Are

	Participant Name	Short Name	Country Code	Logo
1	ENGINEERING – INGEGNERIA INFORMATICA SPA	**ENG**	IT	
2	NATIONAL TECHNICAL UNIVERSITY OF ATHENS	**NTUA**	GR	
3	RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN	**RWTH**	DE	
4	EUROPEAN DYNAMICS LUXEMBOURG SA	**ED**	LU	
5	INTERNATIONAL DATA SPACES EV	**IDSA**	DE	
6	EUROPEAN NETWORK OF TRANSMISSION SYSTEM OPERATORS FOR ELECTRICITY AISBL	**ENTSO-E**	BE	
7	PANEPISTIMIO DYTIKIS ATTIKIS	**UNIWA**	GR	
8	ATOS SPAIN SA	**ATOS**	ES	
9	FUNDACION CARTIF	**CARTIF**	ES	
10	UNIVERZA V LJUBLJANI	**UNILJ**	SL	
11	ENEL X SRL	**ENELX**	IT	
12	REN - REDE ELECTRICA NACIONAL SA	**REN**	PT	
13	CENTRO DE INVESTIGACAO EM ENERGIA REN - STATE GRID SA	**RDN**	PT	
14	UNINOVA-INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIASASSOCIACAO	**UNINOVA**	PT	
15	ENERCOUTIM - ASSOCIACAO EMPRESARIALDE ENERGIA SOLAR DE ALCOUTIM	**ENERC**	PT	
16	FIWARE FOUNDATION EV	**FIWARE**	DE	
17	CENTRICA BUSINESS SOLUTIONS BELGIUM	**CENTRICA**	BE	
18	NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO	**TNO**	NL	

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	Participant Name	Short Name	Country Code	Logo
19	ASM TERNI SPA	**ASM**	IT	
20	VIDES INVESTICIJU FONDS SIA	**LEIF**	LV	
21	COMSENSUS, KOMUNIKACIJE IN SENZORIKA, DOO	**COMSENSUS**	SL	
22	HOLISTIC IKE	**HOLISTIC**	GR	
23	INTERUNIVERSITAIR MICRO-ELECTRONICA CENTRUM	**IMEC**	BE	
24	TERRASIGNA SRL	**TS**	RO	
25	UBIMET GMBH	**UBIMET**	AT	
26	ELEKTRO LJUBLJANA PODJETJE ZADISTRIBUCIJO ELEKTRICNE ENERGIJE D.D.	**EKL**	SL	
27	BORZEN, OPERATER TRGA Z ELEKTRIKO, D.O.O.	**BORZEN**	SL	
28	AJUNTAMIENTO DE SANT CUGAT DEL VALLES	**AJSCV**	ES	
29	ELES, D.O.O., SISTEMSKI OPERATER PRENOSNEGA ELEKTROENERGETSKEGA OMREŽJA	**ELES**	SL	
30	E-LEX - STUDIO LEGALE	**ELEX**	IT	
31	OSMANGAZI ELEKTRIK DAGITIM ANONIM SIRKETI	**OEDAS**	TR	
32	VEOLIA SERVICIOS LECAM SOCIEDAD ANONIMA UNIPERSONAL	**VEOLIA**	ES	
33	STICHTING EG	**EGI**	NL	
34	CINTECH SOLUTIONS LTD	**CN**	CY	
35	EMOTION SRL	**EMOT**	IT	

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Contents

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Figures

Tables

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

History of Changes

Date and Version	Relevant Section	Description
V1.0 01/12/2022	Section 1.1	List of acronyms was updated.
V1.0 01/12/2022	Whole document	The term 'Vocabulary Provider' was replaced by 'Vocabulary Hub' throughout the document.
V1.0 01/12/2022	Section 1.2	Updated reference to Figure 1.
V1.0 01/12/2022	Section 1.4	Reference moved to the end of the first paragraph.
V1.0 01/12/2022	Section 1.4.1	Updated reference to Figure 2.
V1.0 01/12/2022	Section 1.4.1	Updated reference to Figure 3.
V1.0 01/12/2022	Section 1.4.2	Reference updated.
V1.0 01/12/2022	Section 1.4.2	Minor textual changes.
V1.0 01/12/2022	Section 1.5	List of relevant EU projects was extended.
V1.0 01/12/2022	Section 1.5.1	Updated reference to Figure 5.
V1.0 01/12/2022	Section 1.5.2	Minor textural changes.
V1.0 01/12/2022	Section 1.5.3	Reference updated.
V1.0 01/12/2022	Section 1.5.3	Updated reference to Figure 6.
V1.0 01/12/2022	Section 1.5.4	Section added.
V1.0 01/12/2022	Section 1.5.5	Section added.
V1.0 01/12/2022	Section 1.5.6	Section added.
V1.0 01/12/2022	Section 1.5.7	Section added.
V1.0 01/12/2022	Section 1.6	Minor textural changes.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

V1.0 01/12/2022	Section 1.7	Minor textural changes.
V1.0 01/12/2022	Section 2	Added relation to section 1.5.
V1.0 01/12/2022	Section 4.1	Updated figure- and table references.
V1.0 01/12/2022	Section 5.1.1.2	Updated reference to Table 6.
V1.0 01/12/2022	Section 6	Updated reference to Table 7.
V1.0 01/12/2022	Section 7.2	Updated reference to Table 8.
V1.0 01/12/2022	References	List of references was extended.
V1.0 01/12/2022	References	Format was updated.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Executive Summary

Task 4.2 of the BD4NRG project is entitled 'Semantic Business and Platform Interoperability Management'. It deals with the technological implications of the architecture specifications (task T2.4) to enable the discovery and interoperability of the various artefacts that reside within the BD4NRG environment for its participants, e.g. Data, Services, Edge Resources and Machine Learning models, and incorporate meta-data definitions and APIs based on semantic technology.

More specifically, BD4NRG Task 4.2 will provide the necessary functionalities to enable users of data analytics services to join the BD4NRG platform and search and discover services that match their business goals and -needs. In a similar manner, BD4NRG Task 4.2 will provide the means for data analytics service providers to register and publish their services and make them discoverable through semantic meta-data descriptions.

A framework for orchestrating semantic interoperability, based on the International Data Spaces (IDS) initiative, will be developed in WP4 to both register and expose, as well as discover and access data services and data analytics services.

D4.2 is the first report of BD4NRG's Task 4.2. It has addressed the different types of semantic building blocks that are required in the subsequent phases of semantic management of the various artefacts within a federated and open data space approach, as pursued by BD4NRG. These building blocks are meta-data brokering, providing vocabularies, data transformation and data validation.

BD4NRG Task 4.2 will develop the meta-data broker and the vocabulary hub. The architecture and implementation of the meta-data broker and the vocabulary hub will be developed and reported in the follow-up deliverables of BD4NRG Task 4.2, i.e. D4.3 'Architecture document on the integration of selected existing methods, tools, platforms in the BD4NRG platform', to be delivered in M20, D4.5 'An implementation of semantic interoperable components in the BD4NRG platform for selected use cases in the energy domain', to be delivered in M27, and Task 4.2's input for the deliverables of adjacent BD4NRG tasks, i.e. D4.7, D4.8 'BD4NRG Processing / Analytics - Technology Releases', to be delivered in M21 and M28, respectively.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1 Introduction

1.1 List of acronyms

Table 1 contains the acronyms that are used throughout the current document.

Table 1: List of acronyms used in the document.

AI	Artificial Intelligence
API	Application Programming Interface
BD4NRG	Big Data For Energy
CSV	Comma Separated Values
CEN/CENELEC	Comité Européen de Normalisation Électrotechnique (European Committee for Electrotechnical Standardization)
CIM	Core Information Model
D*.*	Deliverable with number, e.g. D4.2
DCAT	Data Catalog Vocabulary
DCAT-AP	DCAT Application Profile
DIH	Data Driven Innovation Hub
DS	Data Space
DSO	Distribution System Operators
EMS	Energy Management Service
EU	European Union
ETSI	European Telecommunications Standards Institute
GCP	Google Cloud Platform
HDFS	HADOOP Distributed File System

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

IDS	International Data Spaces
IDS-IM	IDS Information Model
IoT	Internet of Things
IT	Information Technology
JSON-LD	JSON Linked Data
LSP	Large Scale Pilot
OAS	OpenAPI Specification
OCL	Object Constraint Language
OEMeta-data	Open Energy Meta-data
OEP	Open Energy Platform
OWL	Web Ontology Language
PV	Photovoltaics
Q&A	Question and Answer
QoS	Quality of Service
RAM	Reference Architecture Model
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
REST	Representational State Transfer
RML	RDF Mapping Language
SAREF	Smart Applications Reference Ontology
SCADA	Supervisory Control and Data Acquisition
SHACL	Shapes Constraint Language
SOAP	Simple Object Access Protocol
SWG4	Sub-working group 4

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

T*.*	Task with number, e.g. T4.2
UML	Unified Modeling Language
WP*	Work package with number, e.g. WP4
XML	Extensible Markup Language
XSD	XML Schema Definition
XSLT	Extensible Stylesheet Language Transformations

1.2 BD4NRG Task 4.2

Task 4.2 of the BD4NRG project is entitled 'Semantic Business and Platform Interoperability Management'. It deals with the technological implications of the architecture specifications (task T2.4) to enable the discovery and interoperability of the various artefacts which reside in the BD4NRG environment for its participants, such as Data, Services, Edge Resources and Machine Learning models. Meta-data definitions and APIs are used that are based on semantic technology.

BD4NRG Task 4.2 will provide the necessary functionalities to support data analytics service users to join the BD4NRG platform and to search for (and discover) services that match their goals and business needs. In a similar manner, BD4NRG Task 4.2 will provide the means for data analytics service providers to make use of interoperable data formats when developing their services, as well as register, publish and make them discoverable through semantic meta-data descriptions.

A framework for orchestrating semantic interoperability will be developed to register, expose, and access data services and data analytics services, based on the International Data Spaces (IDS) initiative. IDS is currently gaining major international traction for realizing federated and interoperable data spaces. The IDS Reference Architecture Model (RAM) version 3.0 [9] provides the fundament to develop interoperable data spaces. It aligns very well with the 12 building blocks in the soft infrastructure stack, as defined by the EU OPEN DEI initiative [10] in its report on design principles for data spaces [11].

1.3 BD4NRG Deliverable D4.2

D4.2, *'Report on Existing Methods, Tools and Prototype Implementations to realize the Semantic Interoperability Toolbox, Framework and Platform'* identifies the different types of semantic tools, or building blocks, that are required in the subsequent phases of semantic management of the various artefacts in a federated and open data space approach, as pursued by BD4NRG. These include four types of building blocks: meta-data brokering, vocabulary providing, data transformation, and data validation.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

D4.2 is the first of three deliverables in BD4NRG Task 4.2 and sets the scope and approach for follow-up deliverables:

- D4.3 Architecture document on the integration of selected existing methods, tools, platforms in the BD4NRG platform (M20)
- D4.5 An implementation of semantic interoperable components in the BD4NRG platform for selected use cases in the energy domain (M27)

D4.2 includes a list of existing technologies, in terms of methods, tools and prototypes, that could be considered for the architecture of the semantic building blocks, but does not make a final selection of the tools that are to be used.

1.4 BD4NRG D4.2 in the context of BD4NRG deliverables

Figure 1 shows the BD4NRG Deliverable D4.2 in relation to other Task 4.2 deliverables and adjacent WP2 and WP4 project deliverables.

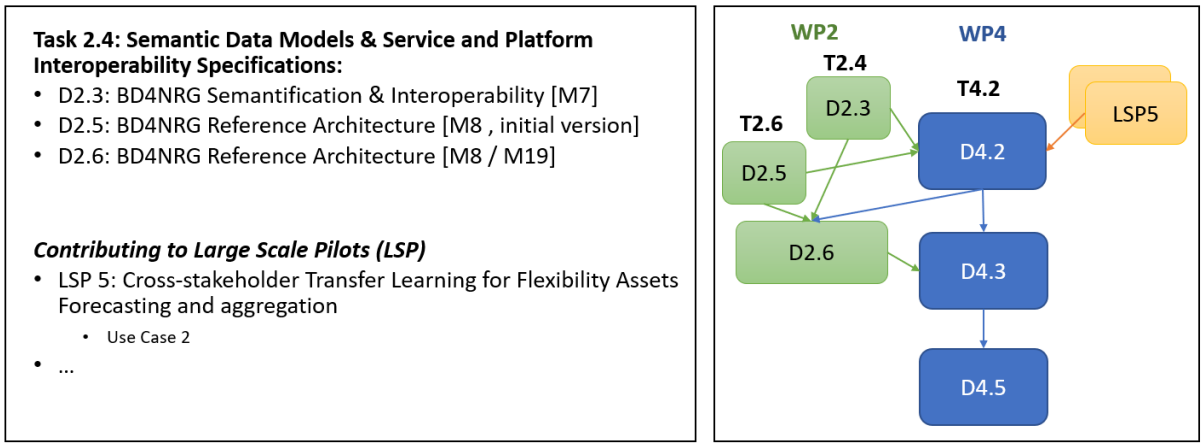

Figure 1: BD4NRG Deliverable D4.2 in relation to adjacent BD4NRG project deliverables.

D4.2 uses the BD4NRG WP2 deliverables D2.3 and D2.5/D2.6 as input. The input of these deliverables is summarized in the following paragraphs.

1.4.1 Input from deliverable BD4NRG deliverable D2.3: summary

BD4NRG deliverable D2.3 describes the requirements for interoperability of data spaces in the context of BD4NRG. The decentralized and distributed nature of data spaces has important implications for interoperability specifications. Data providers must be able to publish and describe their data resources in such a way that they become discoverable for (potential) data consumers, whoever they may be. The core of any data space interoperability specification must therefore contain an

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

unambiguous meta-data model that is understood and agreed upon by every participant in the data space [7].

Based on D2.3, the first thing that is required for BD4NRG is a meta-data repository -or broker- in which data resources are published so that these resources are discoverable for consumers. More specific, the meta-data broker must implement the meta-data categories as described in D2.3 chapter 2 (Figure 2): content, context, and concept [7].

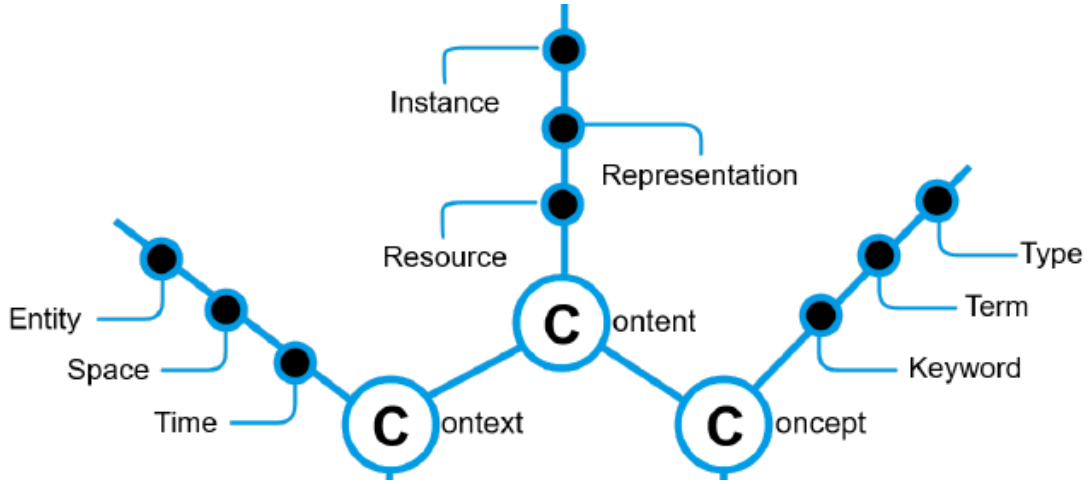

Figure 2: Content, Concept and Context (see D2.3, Chapter 2) [9].

Content contains the most basic elements of meta-data, e.g. file size, file format, and file creation. Context contains meta-data relating to temporal and spatial aspects, e.g. what time period the data cover, when and where they was gathered, and whether they belong- or relate to a larger dataset or entity. Finally, concept meta-data describes the actual data and how they should be interpreted, e.g. what type of observation the data refer to, what kind of objects they represent, and the meanings of certain data parameters.

The second thing that is required for BD4NRG (based on D2.3) are vocabularies that are fit for purpose to BD4NRG. Examples of these vocabularies have been described in D2.3, chapter 4, and are based on the data clustering given in D2.3 chapter 3 [7] that resulted from an analysis of the data to be used in the large scale pilots (LSPs).

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

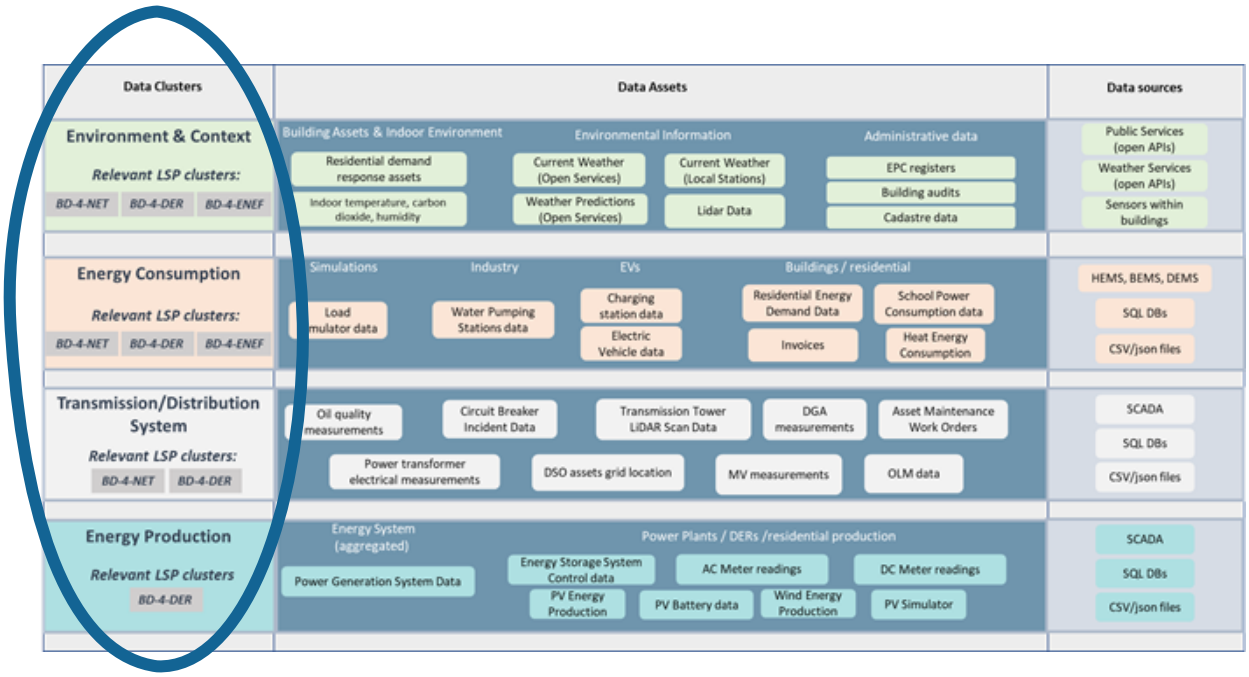

Figure 3: Overview of Data Clusters relevant for BD4NRG. The project LSPs have identified a number of data types from corresponding data sources that will be used by the analytics engines in order to create value.

Figure 3 shows an overview of Data Clusters that are relevant for BD4NRG. These clusters are briefly described below.

Energy Production data are provided as aggregated data for the energy system or power plants, distributed energy resources, and residential energy production systems. These also include data from renewable energy sources such as PVs and wind turbines. Such data can be extracted by SCADA systems and are stored inside (SQL) databases or files (e.g. as CSV or JSON). Energy production data are relevant for the BD-4-DER (LSP5, LSP6, LSP7, LSP8, LSP9) cluster.

Data coming from the Transmission / Distribution System mainly concern data relating to the different assets of the grid. These include data about power transformers and transmission towers, the location of these assets, related incidents, and workorders for maintenance. Similarly, to Energy Production data, the data related to the Transmission / Distribution System can be extracted by SCADA systems and is stored inside databases or CSV or JSON files. Transmission and Distribution System data is relevant for the BD-4-NET (see LSP1, LSP2) and BD-4-DER (see LSP4, LSP5, LSP6) clusters.

Energy Consumption data is quite diverse. They relate to building- and residential energy consumption, charging of electric vehicles, large industrial power demands and simulation data. The data is provided by different energy management systems (such as HEMs, BEMs, DEMs) or are stored in databases or files. Energy consumption data is relevant for all pilot clusters: BD-4-NET (LSP3); BD-4-DER (LSP6, LSP8, LSP9); BD-4-ENEF (LSP10, LSP11, LSP12).

Finally, Environment and Context data will be used in the BD4NRG analytics services. Such data is related to i) building assets (e.g. appliances that can be used in demand-response scenarios), ii) measurements for indoor environments, iii) environmental information, mainly focused on local- and

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

area wide weather data, and iv) administrative data from public services using open APIs that is provided by the data owners (e.g., weather services or public administration services). Environment and Context data are relevant for all pilot clusters in BD4NRG. Environmental information is provided and used in BD-4-NET (LSP3, LSP4); BD-4-DER (LSP6, LSP7, LSP9); BD-4-ENEF (LSP10, LSP11, LSP12), data related to building assets and indoor environments are provided and used in BD-4-DER (LSP5); BD-4-ENEF (LSP10), and public administration related data are provided and used in BD-4-ENEF (LSP10, LSP11).

1.4.2 Input from deliverable BD4NRG deliverable D2.5: summary

BD4NRG deliverable 2.5 [8] offers a first version of the BD4NRG reference architecture based on two analyses:

- D2.5 integrates the different requirements and goals of the large-scale pilots into one picture. It does so by mapping of the use case descriptions from the Large Scale Pilots to the BRDIGE RA model.
- D2.5 analyses existing Data Space project coalitions and efforts, notably IDSA, GAIA-X and FIWARE.

The BD4NRG RA that is proposed in D2.5 takes the BRIDGE RA and complements it by a vertical pillar, which is concerned with the integration of Data Space enablers.

D2.5 makes the following choices in the BD4NRG RA that are relevant for D4.2:

- For the BD4NRG-Analytics Services layer, D2.5 selects the ALIDA Big Data analytics platform as a starting point. In the executive summary of D2.5 it is noted that "WP 4 will adopt and extend ALIDA during the project."
- The BD4NRG-Data Governance layer will build upon the integration of IDSA, GAIA-X and FIWARE. The conclusion section of D2.5 notes that "[these initiatives] are of great importance for the project and will be integrated in the work of the technical Work Packages 3-5."

Finally, D2.5 provides implementation recommendations (i.e. existing methods, tools and prototypes) based on a review of Linux Foundation Energy projects. We used these as criteria to evaluate existing methods, tools and prototypes:

- Use microservice approach as much as possible.
- Use of REST and GraphQL are recommended.
- Use of SOAP is not recommended.
- Aim for intermediate data representations, such as with NGSI-LD, to reduce # of mappings.
- Tools should be secure, covering authorization as well as defense against malicious attacks.
- Documentation should be good, i.e. up-to-date, well-structured and clearly written.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1.5 Relevant EU projects

BD4NRG Task 4.2 will build and extend upon the work done in related and relevant EU projects. As such, the following paragraphs in this section briefly describe the work of the following related EU projects in relation to the work of T4.2:

- the EU Interconnect project,
- the EU AI4EU project,
- the EU EUHubs4Data project,
- the EU BIG IoT project,
- the EU INTER-IoT project,
- the EU symbIoTe project, and
- the VICINITY project.

1.5.1 INTERCONNECT

The EU InterConnect project[1] delivers interoperable solutions/services connecting (devices in) Smart Homes, Buildings and Grids for the democratization of efficient energy management, through a flexible and interoperable ecosystem where demand side flexibility can be soundly integrated with effective benefits to end-users.

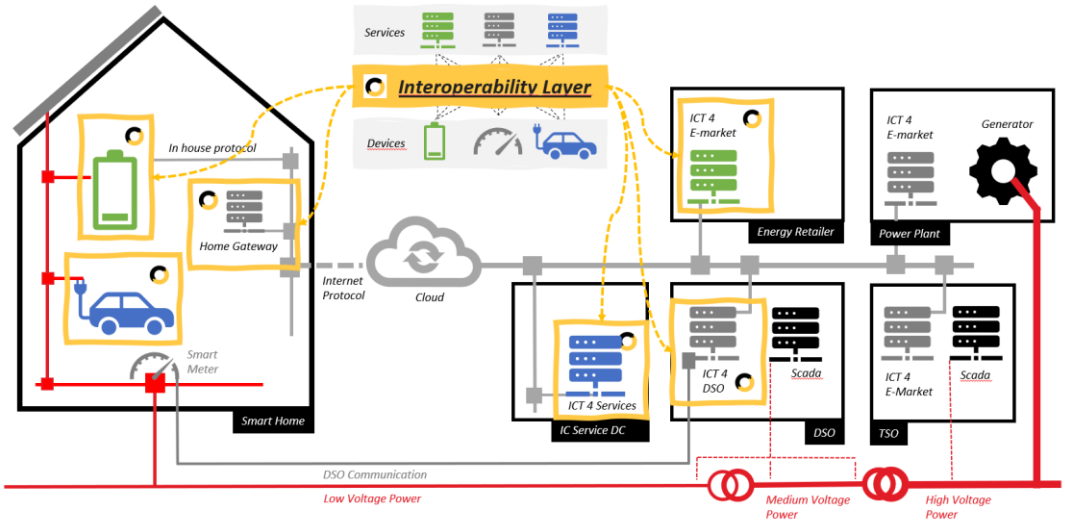

Figure 4: InterConnect overview of components and the Interoperability Layer.

The concept of semantic interoperability plays a pivotal role in the creation of an ecosystem of distributed and interoperable components, connected through the Internet and Smart Grids. This

[1] https://interconnectproject.eu/

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

ecosystem spans across different domains (of expertise), like Energy and Internet of Things. The ecosystem also has many types of stakeholders: device users and -owners, facility managers and inhabitants of buildings, device manufacturers, IoT platform providers, energy service providers, electricity providers and -retailers, distribution system operators (DSO), etc.

The InterConnect project assumes the previous existence of architectures, information models, development frameworks and other standards in the domains it spans across. Instead of trying to replace these existing (also technological) agreements between stakeholders with InterConnect versions, the idea of an 'Interoperability Layer' is applied. This is visualized in Figure 4, where devices (a Home Battery and Electric Vehicle) in green, blue, and grey colors are wrapped inside a yellow-colored interoperability layer from InterConnect, just like on-line services (implemented as web- or Internet applications in Datacenters). This allows the devices to interact with a local energy management service (EMS) on a Home Gateway, and distributed EMS across energy retailers and DSOs that are connected by the Smart Grid.

Determining where to introduce ecosystem component boundaries is done by using the design principle of separation of concerns. The less an ecosystem component needs to know about the (internal) workings and the context of another domain's component it interacts with, the better.

To keep interoperability at a high level of abstraction, wherever possible the InterConnect project models (component) interaction as the exchange of knowledge. Devices and services are seen as Knowledge Bases that exchange knowledge, that is encoded in transmissible data using semantic web technology. The InterConnect project has created a set of ontologies that describe the shared understanding of semantical concepts across the domains that are involved in the InterConnect ecosystem. These concepts are a combination of concepts in existing ontologies (e.g. SAREF from ETSI), existing information models (e.g. from CEN/CENELEC), and newly created concepts and relations in InterConnect. The latter act as a conceptual bridge between different domains (IoT, Energy) of expertise. The project has also tried to stay aligned with existing standards as much as possible, as to stimulate uptake by the industry of the Interoperability Layer to interconnect existing domains.

Using the InterConnect set of ontologies it is possible to create knowledge graphs that can be exchanged between Knowledge Bases with the help of the (InterConnect) Interoperability Framework, which – next to a Service Store with access control - provides a Generic Adapter software implementation. This adapter allows ecosystem components to connect to the (semantic) Interoperability Layer. An important part of the implementation is the so called 'Knowledge Engine', which can process knowledge graphs and perform reasoning both based on the knowledge contained within these graphs and that in its corresponding Knowledge Base. Note that, at the moment of writing, the project is still in progress, as are its results.

Seven large pilots across Europe will be used to get practical feedback on the use of the ontologies, and then a potentially revised set can be used as input for ontology standardization, making it even more attractive for InterConnect ecosystem to use these ontologies for interoperability.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1.5.2 AI4EU

AI4EU [1] was a Europe-wide artificial intelligence (AI) project that develops an AI experimentation platform facilitating collective work in AI research. The aim of the project was to lower the barriers for AI innovation in the EU, while boosting technology transfer and related business development. The AI4EU platform acts as an AI services broker and provides a one-stop shop for services, algorithms, software frameworks, development tools, components, modules, data, computing resources, prototyping functions, and access to funding.

The AI4EU Experiments platform offers several services for the creation of human-centered AI-solutions, building modular structures and using hybrid AI technologies. Services and data Interoperability is one of the main project objectives ensuring "out of the box" use of the services and data. Essentially the AI4EU platform provides the following functionalities:

- visual composition of AI pipelines,
- the ability to make use of trained models with published and well-known interfaces,
- functionalities to connect data sets via brokers or data streams,
- a marketplace allowing service developers and data providers to publish tools, data resources or solutions,
- a collaboration environment supporting the creation of teams that collaborate on the development of AI pipelines.

The AI services hosted in the AI4EU platform are not specific to -or targeted at- a particular domain and implement generic AI models for tasks such as classification, regression, data driven predictions, and data transformations. Although there is no specific focus on the energy domain there is at least one energy related service provided in the current AI4EU marketplace [2] that addresses the problem of predicting electricity demand of a boiler room in a large District Heating Network.

The AI4EU platform and interoperability approach is based on the Acumos [3] open-source framework for building, sharing, and deploying AI apps. Acumos is a project of the Linux AI & Data Foundation that supports open-source innovation in artificial intelligence, machine learning, deep learning, and data. Figure 5 depicts an overview of the four stages of AI development that are supported by Acumos, while the framework provides specifications for data interoperability, service interfaces interoperability, as well as data models for describing the available AI services and including them in the Acumos marketplace.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

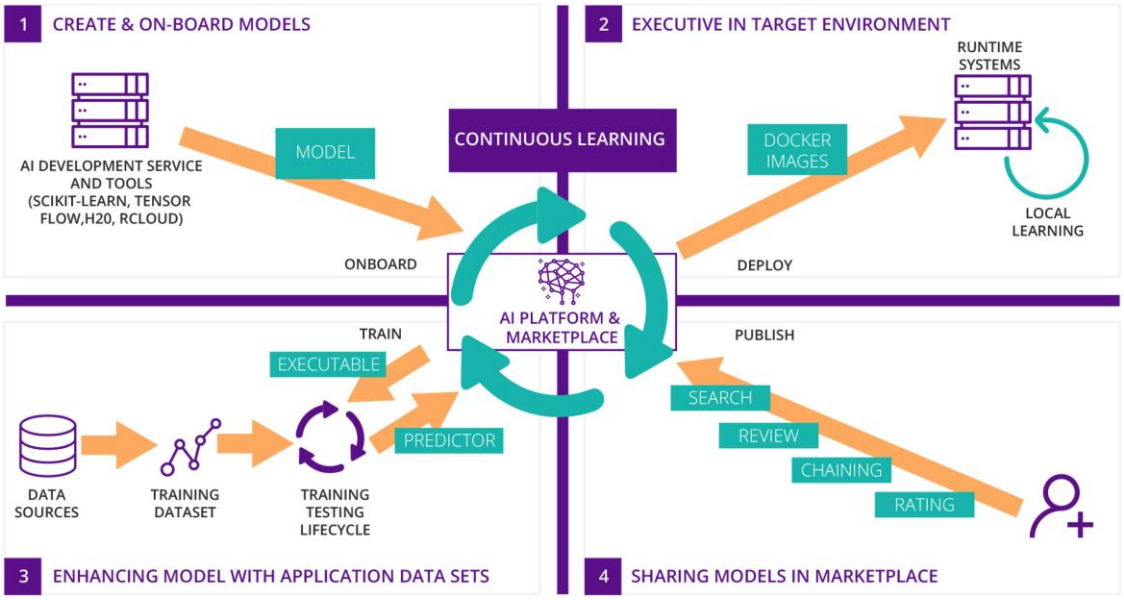

Figure 5: The AI development stages supported by Acumos[2].

1.5.3 EUHubs4Data

The EUHubs4Data project is part of the European Federation of Data Driven Innovation Hubs (DIHs). Its aim is to service as a reference to the establishment of the Common European Data Spaces. The EUHubs4Data project realizes several components to foster data interoperability, one of them being a federated catalogue for datasets and services. The federated catalogue of the EUHubs4Data project is accessible via a web browser [4].

WP4 of the EUHubs4Data project is of main interest for BD4NRG. EUHubs4Data project deliverable D4.4 [5] describes interoperability requirements based on the International Data Space (IDS) Reference Architecture Model (RAM). This includes the functional requirements for a federated catalogue, thereby distinguishing:

- the catalogue function, i.e., how resources can be registered and exposed and made findable for users, and
- the federation function, i.e., how individual catalogues for various data space can be federated into a 'virtually single' catalogue.

Figure 6 shows the ambition of the federated catalogue as pursued in the EUHubs4Data project, where data spaces (DS) A, B and C are connected to form a federation and their individual catalogues are interconnected to act as a single, overarching catalogue. This allows service publication and -discovery across data spaces.

[2] https://www.acumos.org/platform/

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

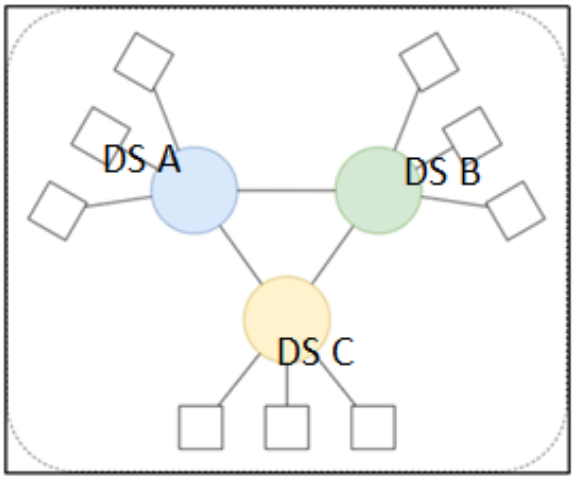

Figure 6: Scenario in which hubs are part of a federated network. Conceptually the network of hubs can act as a federated catalogue [5].

1.5.4 BIG IoT

The BIG-IoT project (Bridging the Interoperability Gap of the IoT)[3] was a project within the ICT-30-2015 call. The project addressed interoperability in an Internet of Things (IoT) ecosystem. The aim of the project was to enable the emergence of cross-platform, cross-standard, and cross-domain IoT services and applications toward building IoT ecosystems, in order to connect service and 'thing' providers to users, by leveraging Semantic Web technologies [13].

Figure 7 shows the BIG IoT core architecture concepts and the interactions between them. Offerings encompass a set of IoT resources (e.g. sensor data) that are registered in the BIG IoT marketplace. Providers register their offerings via a common API. Consumer discover (and are able to subscribe to) offerings of interest via the marketplace so that they can access them. An offering is registered using an offering description. This description contains meta-data about the kind of data and how it can be accessed, and is provided in a machine-interpretable manner using RDF models [12]. The offerings descriptions use common (shared) information models (e.g. using the Semantic Web and linked data), which support providers to describe their resources offerings so that consumers can find and interpret them [13]. Discovery of offerings by consumers is done using queries that entail specifications of the type of offerings (e.g. the type of data, maximum price, desired license types, region, etc.) [12].

[3] https://cordis.europa.eu/project/id/688038

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

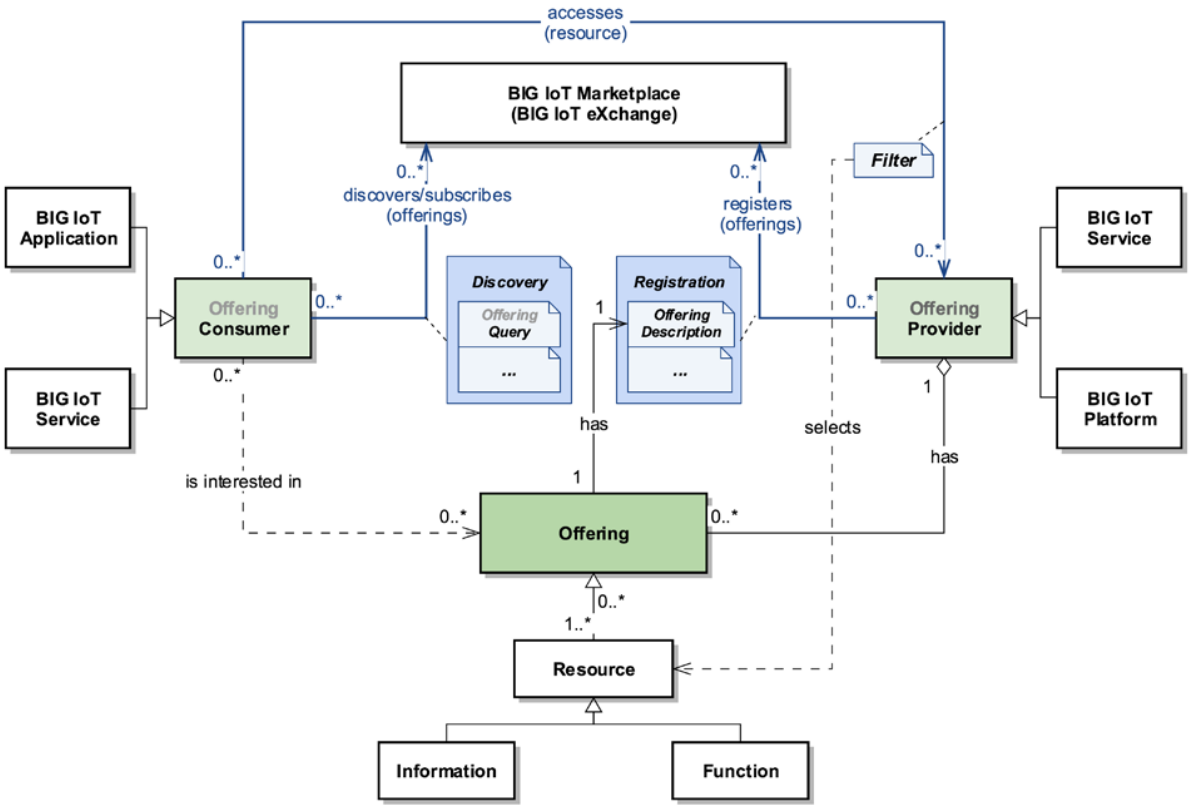

Figure 7: BIG IoT Core architecture concepts and terminology, namely Offerings, Offering Providers, Offering Consumers, Queries and Filters [12].

1.5.5 INTER-IoT

The INTER-IoT[4] project was an ICT-30-2015 project that aimed is to design and implement a framework that allowed interoperability among different IoT platforms [14]. One of the project's key beliefs was that interoperability can be addressed on multiple levels in the software stack, but that the key to interoperability is the use of common description- and data representation frameworks that define 'things', capabilities and data in machine-readable and -interpretable forms, in that ontologies are used for semantic annotation (resource- and offering descriptions) and resource- and offering discovery [15]. The INTER-IoT solution for interoperability was layer-oriented to provide interoperability at any layer and across layers among different IoT systems and platforms. These layers were [18]:

- At the Device layer: inclusion of new IoT devices.
- The Networking layer: support for smart objects mobility (roaming) and information routing.
- The Middleware layer: resource discovery and management.
- The Application and Services layer: the discovery, use, import, export and combination of heterogeneous services between different IoT platforms.

[4] https://cordis.europa.eu/project/id/687283

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- The Data and Semantics layer: a common interpretation of data and information, providing semantic interoperability.

In the context of the current deliverable, we focus on the functionalities of the Application and Services layer (AS2AS) and the Data and Semantics layer (DS2DS).

AS2AS in INTER-IoT was addressed using a Service Catalogue approach, so that services and applications could be registered by providers and discovered by consumers. The identified functionalities for a Service Catalogue and Service Discovery were [18]:

- register services and applications to make them discoverable,
- offer a description or detailed information about the services and applications,
- use common metadata annotations in order to create a point of interoperability,
- allow publishing linked-data descriptions as meta-data of resources,
- unify data catalogue with semantics,
- discover information about (IoT) services.

DS2DS addressed understanding of structure and meaning of data using meta-data descriptions, to be used for different purposes, such as resource discovery (combined with the Service Catalogue), resource management, and access control. The DS2DS solution can be used for semantic translation of data received e.g. from IoT applications and services in the AS2AS layer, and semantic annotations can be used in the Service Catalogue to enable service description and discovery [18].

1.5.6 symbIoTe

The H2020 project symbIoTe[5] (ICT-30-2015 call) objective was "to create a mediation framework to enable the discovery and sharing of connected devices across existing and future IoT platforms to enable platform federation and rapid development of cross-platform IoT applications." [16]. Regarding semantic interoperability the project had an approach of an common ontology with basic concepts for all platforms connected via the symbIoTe framework that can be extended with domain and platform specific concepts. The basic concepts are sufficient to enable generic interoperable mediation service. The extensions are needed for semantic and syntactic transformation services.

The project used the Semantic Web technology stack (ontologies, SPARQL, RDF) and addressed the topic and difficulties of semantic mappings and alignments to resolve ontological mismatches. Figure 8 shows a schematic representation of semantic mapping, how it can be used for semantic interoperability and which kind of software and tools are involved. In this approach to semantic interoperability, the SPARQL Query Re-Writing method is used to define the mapping and execute the semantic mediation.

[5] https://cordis.europa.eu/project/id/688156

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

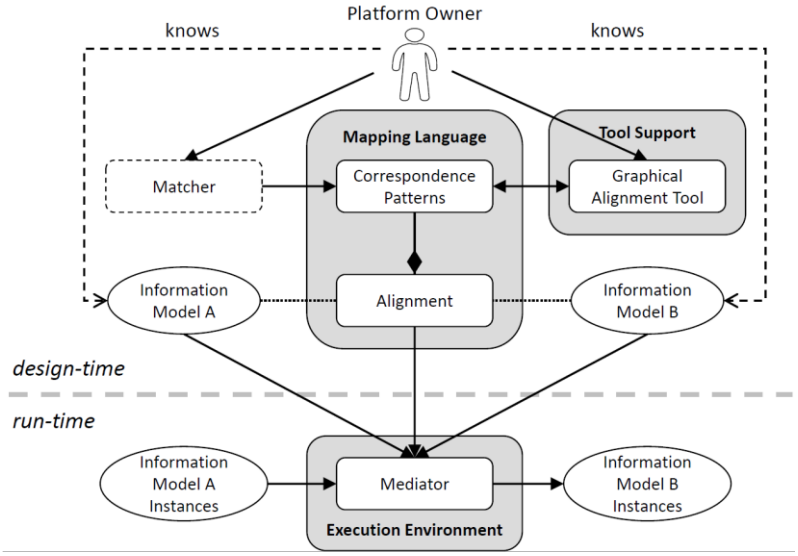

Figure 8: Schematic representation of semantic mapping [16].

In the architecture of symbIoTe, semantic interoperability is achieved via a generic mediation service, which was delivered by the project. Different IoT platforms and solutions are not required to deal with interoperability complexity themselves, but can delegate it to a shared service where mappings are defined using an extendable core information model (CIM). Figure 9 depicts the high-level diagram of the symbIoTe approach.

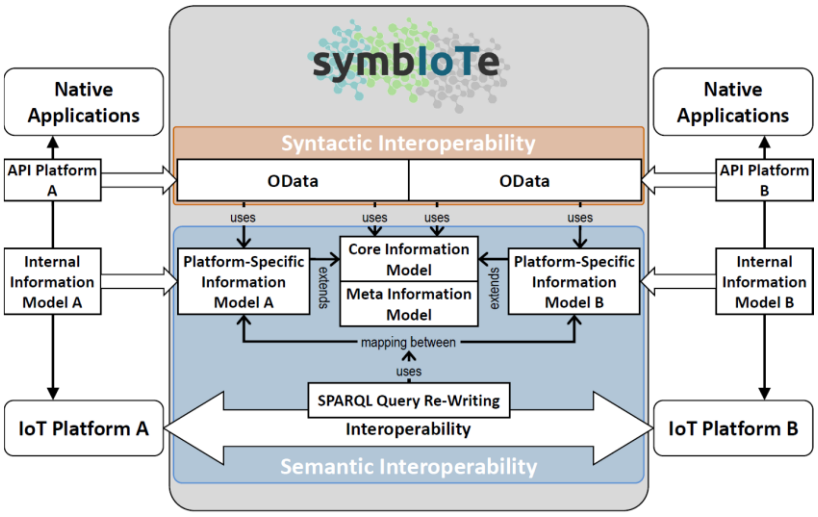

Figure 9: High-level diagram of symbIoTe approach for semantic interoperability [16].

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1.5.7 VICINITY

The H2020 project VICINITY[6] (ICT-30-2015 call) addressed interoperability for Internet of Things (IoT) ecosystems by providing a decentralized platform. The project identified lack of consensus in IoT standards that hinder the interoperability among IoT ecosystems, including technical and semantic interoperability. VICINITY implemented "an open virtual neighborhood to interconnect IoT ecosystems and smart objects providing transparent interoperability based on Web of Things." [17].

The described Web of Things approach in VICINITY included submitting a *Thing Description* and *Thing Ecosystem Description* to dynamically discover new ecosystems. The *Thing Description* contains semantic metadata that explicitly specifies the semantics used by an IoT device and how to interact with the device. Other components in the VICINITY approach included *Nodes*, *P2P Network*, *Neighborhood Manager, Semantic Agent Platform*, *Gateway API* and *Ontology Network*.

At first glance the approach looks similar to (and pre-dates) the INTERCONNECT project architecture (see section 1.5.1). In both architectures key concepts are:

- the Semantic Web technology stack (i.e. the use of ontologies and SPARQL),
- having a common information model (SAREF-ontology for INTERCONNECT, WoT-ontology for VICINITY), and
- a service discovery protocol in a decentralized platform architecture.

1.6 Scoping semantic interoperability

Semantic interoperability enables people and organizations to understand each other and IT-systems to be interoperable. It consists of domain knowledge, shared (business) language, terminology, definitions, data models and meta data that are made available in formal machine-readable format. These machine-readable specifications include OWL-ontologies and UML-class diagrams for data models, OpenAPI specifications and XML/JSON schemas for data exchange, and OCL, SHACL or Schematron for business rules and -constraints specifications.

Here, three functional areas are distinguished to support the lifecycle and usage of semantic specifications:

- Knowledge engineering – the practice of making domain knowledge explicit. This includes the creation of (shared) data models, ontology engineering, annotating data sets and data services, and terminology definitions and mappings.
- Semantics governance – the practice of sharing, maintaining, and governing semantic specifications. This includes cataloguing of specifications, providing usage support, collaborative development and -adaptation of specifications according to changing requirements, version control, publishing releases, and many other aspects that relate to governing semantic specifications.

[6] https://cordis.europa.eu/project/id/688467

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- Semantic IT configuration – the practice of using semantic specification to generate or configure (parts of the) software automatically. For instance, for REST APIs there is the OpenAPI Specification (OAS) standard to specify APIs in a machine-readable format, SwaggerHub to share and maintain these specifications in a collaborative way, and code generators that are able to use the OAS as input to automatically create client- and server software code.

1.7 Structure of this report

Chapter 2 describes the required building blocks for Semantic Interoperability in BD4NRG. These building blocks are the result of the analysis performed in deliverable BD4NRG deliverable D2.3 [7]. Chapter 2 provides a short summary of deliverable D2.3.

Chapters 3, 4, 5 and 6 address existing methods, tools, and prototypes for each Semantic Interoperability building block that are both relevant and useful, and *could be considered* for the architecture and implementation within the context of BD4NRG and the large-scale pilots (LSPs). Please note that the current deliverable D4.2 does not contain a final selection of the tools that are to be used.

Finally, chapter 7 summarizes the findings of chapters 3, 4, 5 and 6 in an overall conclusion, that is, which building blocks are of value for BD4NRG and the LSPs, which of these building blocks are reasonably readily available, and which of them are missing and should be developed within the scope of BD4NRG.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

2 Semantic interoperability building blocks

Based on the relevant EU projects (see section 1.5) distributed and open ecosystems such as BD4NRG are heterogeneous in nature, where different actors have different goals and interests and use different tools and knowledge levels to achieve them. Semantic interoperability addresses the heterogenic nature of the ecosystem by ensuring that meaning and understanding of data are preserved across actors. It is additional to the technical level of interoperability and has to be facilitated by specific components or building blocks with specific functionalities.

First, the data within an ecosystem need to be discoverable. A consumer needs to be able to lookup types of data and services, based on descriptions that are stored as meta-data. In the same fashion, a provider needs to be able to publish (register) these meta-data. This registry role is attributed to a broker or meta-data catalogue.

Next, the data need to be describable, using a collection of concepts, their descriptions, and relations. These collections, i.e., vocabularies, need to be stored in a vocabulary hub. The vocabulary hub enables both the description of data by data providers and understanding of data by data consumers.

Finally, data transformations and data validation building blocks are required in the BD4NRG system, due to the previously mentioned heterogeneity of data models and -sources. The rules for data conversion, -transformation, and -validation can be specified in machine-readable semantic models.

This deliverable D4.2 addresses the supporting building blocks to facilitate semantic interoperability in the BD4NRG architecture. The next chapters further elaborate on each of the building blocks: the broker, the vocabulary hub, and data transformation and -validation. Each chapter provides a functional description of the building block, its relation to BD4NRG, and existing methods, tools, or prototypes that are currently available.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

3 Broker / meta-data catalogue

A catalogue is a searchable list of items with descriptions about their details. Meta-data catalogues allow finding, accessing, and understanding data within an organization, between organizations inside a data space, and even across different dataspaces.

Meta-data catalogues are used by both data providers and -consumers. A data consumer must be able to discover data, have access to the data, and be able to use the data. Regarding the latter, a consumer must have knowledge about the data interface, -structure, and whether the data are actual. It might also be necessary to know more about its origin, ownership and whether it is legally allowed to use those data. A data provider uses a meta-data catalogue to advertise, explain, and govern data.

A meta-data catalogue can be implemented as a federated system: a system that maps multiple autonomous, individual systems into a single federated whole. A federated catalogue is a composition of various catalogues inside a networked system. Its parts are often geographically scattered, while collaborating in such a way that users experience them as one single system.

Within a federated catalogue there is no central integration of meta data. Using a distributed query mechanism, a federated catalogue acts as one single catalogue. A federated catalogue decomposes a search query, posed by a data consumer into subqueries and sends them to the individual catalogue systems (brokers). After processing the search query, the system composites the results back into one result and sends it back to the data consumer.

The role of a federated catalogue in BD4NRG is to enable effective collaboration across sectors and regions by offering a single point of entry to find and use resources, i.e., data resources, processing resources, and service resources.

3.1 DCAT-AP

Many organizations map their available data sources in so-called data catalogues. This includes data that is available both internally and partly publicly for search. To present datasets in an orderly manner and to be able to search for datasets in a targeted manner, datasets are described with meta-data. To avoid forcing users to visit all individual catalogs to find what they need, meta-data from various regional, national and theme catalogs is collected. The DCAT standard has been developed to standardize the exchange of meta-data between catalogs. The interoperability between catalogs allows for one-time entry and multiple use.

The so-called ‘DCAT-application profile’ (DCAT-AP) for datasets is the European specification of the meta-data that European data portals use for the exchange of meta-data about datasets between catalogs. The central European data portal, found at https://data.europa.eu and managed by the EU Publication Office, has been set up based on DCAT-AP.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

3.2 IDS Meta-data broker (implementation)

The IDS Meta-data Broker[7] is a comprehensive connector that provides the necessary interfaces for communicating with any other International Data Spaces connector. More specifically, it is capable of handling messages from IDS connectors, indicating a status update, such as new data being available. Self-descripting meta-data of connectors are automatically indexed and can be optionally restricted by means of usage control policies. Search functionality encompasses a full-text search, filter options based on the International Data Spaces information model, and full queries using SPARQL, the standardized query language in the semantic web. Human user-friendly access is granted via a web interface.

The IDS Meta-data Broker supports every interaction defined in the process layer, the descriptions defined in the information layer, and the architectures listed in the system layer. A meta-data broker can provide additional services when needed. These can then be described as well with machine-readable descriptions compliant with the IDS information model. For example, a broker can regularly execute heartbeats to detect inactive connectors. The storage of quality of service (QoS) metrics or payment models can be implemented as well.

3.3 IDS Information model

The IDS Information Model (IDS-IM)[8] is an RDFS/OWL-ontology covering the fundamental concepts of the IDS[9], i.e., the types of digital contents that are exchanged by participants by means of the IDS infrastructure components. The model development is led by the Fraunhofer Institutes for Applied Information Technology FIT and Intelligent Analysis and Information Systems IAIS with support by members of the International Data Spaces Association in the context of the Information Model sub-working group (SWG4).

Development of the IDS Meta Model is based on GitHub, following a defined branching model. Contributions and community feedback are maintained via the GitHub ticketing system. The release process is aligned with the International Data Spaces Association architecture working group meetings, i.e., there are roughly 2 releases scheduled per year with intermediary updates to the development branch. The current release version is 4.1.0, with the latest revision 4.1.0. The Information Model and associated resources published on GitHub are available under the Apache License 2.0.

3.4 Other meta-data catalogues software

Below is a (non-exhaustive) list of other meta-data catalogue software. The descriptions are based on -or taken from- the product websites, which are mentioned in the footnotes.

[7] www.dataspaces.fraunhofer.de/en/software/broker.html

[8] github.com/International-Data-Spaces-Association/InformationModel

[9] w3id.org/idsa/core

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Amundsen[10] is a data discovery and meta-data engine that indexes data resources (tables, dashboards, streams, etc.) and uses a page-rank style search based on usage patterns (e.g., highly queried tables show up earlier than less queried tables).

DataHub[11] is an open-source meta-data platform developed by LinkedIn. It is built on top of Kafka streams and uses a push-based architecture. This means that data is published on DataHub by other services who push the data. Because of its streaming (Kafka) nature, it allows near-real time updates for data consumers.

Marquez[12] is an open-source meta-data service (released and open sourced by WeWork) that can be used for the collection, aggregation, and visualization meta-data in data ecosystems. It provides insights into how datasets are consumed and produced and provides global visibility into job runtime and frequency of dataset access and -lifecycle management.

Metacat[13] is described as "a unified meta-data exploration API service" that is compatible with Hive, RDS, Teradata, Redshift, S3 and Cassandra. It aims to provide information about the type of data, where it is located and how to process it.

Databook[14] was developed by Uber. Like DataHub it uses a push-based architecture for streaming data (i.e. Kafka messages). Databook was designed to handle large volumes of data (e.g., Kafka messages) that are stored in HDFS across multiple data centers.

Lexikon[15] was developed at Spotiy. It is a non-open source solution that is similar to DataHub and DataBook, with the addition of personalized data set suggestions.

Dataportal was developed by Zeenea[16] and is used by Airbnb. It is a database that supports meta-data models, data profiling (statistics), data lineage (over time), and data discovery.

Apache Atlas[17] is described on its website as "a scalable and extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem." Apache Atlas aims to provide meta-data management and governance, in order to catalog and classify data assets.

Data Catalog[18] is a centralized cloud-based service for meta-data management by Google Cloud. It is to provide an optimized search index for data assets belonging to GCP (Google Cloud Platform) projects, including datasets, tables, views, text/CSV files, spreadsheets, and data streams.

[10] www.amundsen.io
[11] datahubproject.io
[12] github.com/MarquezProject/marquez
[13] github.com/Netflix/metacat
[14] eng.uber.com/databook
[15] engineering.atspotify.com/2020/02/27/how-we-improved-data-discovery-for-data-scientists-at-spotify
[16] zeenea.com/data-catalog
[17] atlas.apache.org
[18] cloud.google.com/data-catalog

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

3.5 Conclusion

A (federated) meta-data catalogue, as described functionally in this chapter, is crucial for the success and scalability of the BD4NRG ecosystem. The BD4NRG ecosystem is there to enable controlled sharing of energy data to enable improved data-driven analytics and new (analytic) services. These data is spread across numerous parties and systems that need to agree upon legal-, organizational-, semantic- and technical interoperability matters in a scalable way. It is not cost efficient nor manageable to make bilateral agreements with every potential organization in the ecosystem. The catalogue functionality allows organization to uniformly publish their offerings and annotate with the relevant interoperability meta data. This allows other organizations to search & find service offerings that match their need.

The catalogue should support the meta-data registration of data services, including:

- Data sets
- Data services / APIs
- Machine learning models (e.g. for Algorithm-to-Data scenarios)
- Other processing resources

Data services are developed in WP3 and can be registered in such a catalogue (WP4) to allow for direct use in the pilots and/or market place functionality (WP5).

The IDS Information Model, the DCAT-standard and DCAT-AP in particular can be used as a starting point for the conceptual model of the catalogue. It needs to be extended for machine learning models and processing resources.

When providing a meta-data catalogue implementation for the BD4NRG ecosystem, further investigation is needed which implementations, mentioned in sections 3.1, 3.2, 3.3 and 3.4 can be reused and what needs to be added or changed to make it fit for purpose.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

4 Vocabulary hub

As semantic models (also called vocabularies) are at the heart of semantic interoperability, these models need to be findable, accessible, and usable for users and services in the BD4NRG ecosystem. The broker / meta-data catalogue as described in the previous section is a catalogue service for data sets, data services and other runtime assets. The vocabulary hub is about catalogue functionality for the design-time models that semantically describe those runtime assets. This includes ontologies, reference data models or meta-data elements that define the data itself, annotate the assets or define the semantic data transformation and validation (see sections below).

The vocabulary hub must at least provide functionality to store and publish vocabularies and enable collaboration. Collaboration may comprise search, selection, matching, updating, request for changes, version management, deletion, knowledge sharing, Q&A and other supporting functions.

Like with the meta-data catalogue, the vocabulary hub could also be federated to enable effective collaboration across sectors and regions by offering a single point of entry to find and use semantic models.

4.1 Smart Data Models

The Smart Data Models[19] initiative is an agile standardization initiative that, among others, provides models related to the BD4NRG, including energy management, production, and distribution, weather, environment, etc. The proposed models are released with an open license allowing the user the free use, free modification, and free sharing of the modifications without any other restriction. Source for the models is available at GitHub[20]. BD4NRG deliverable D2.3 provided an introduction to the Smart Data Models initiative and showed its relevancy to the project. In the context of task T4.2 the related work continued, and a set of relevant data models were identified and mapped to the data clusters of BD4NRG as shown in Figure 3.

The outcome of this work is presented in Table 2, Table 3, Table 4, and Table 5, which aim to support the BD4NRG ecosystem to select the proper data models for semantically enriched data representations. A short description is provided per relevant data model along with a link to the corresponding repository. Note that a data model can be relevant for more than one data clusters. Moreover, the Smart Data Models initiative is evolving, and new models are continuously added. The provided list of relevant data models is not exhaustive and will be evolving throughout the BD4NRG depending on the needs of the data driven applications that are being developed.

[19] https://smartdatamodels.org

[20] https://github.com/smart-data-models

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 2: Smart Data Models in energy production.

Energy Production Smart Data Model	Description and link
Device	Allows to represent devices of different nature (IoT, mobile, wearable, etc.). Sensors/devices in the transmission/distribution system can be represented using this model. https://github.com/smart-data-models/dataModel.Device
DeviceModel	Captures the static properties of a Device. https://github.com/smart-data-models/dataModel.Device/blob/master/DeviceModel/README.md
Substation	Represents equipment through which electric energy in bulk is passed for the purposes of switching or modifying its characteristics. https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/Substation
PowerTransformer	Can be used to represent power transformers: electrical device consisting of two or more coupled windings, with or without a magnetic core, for introducing mutual coupling between electric circuits. https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/PowerTransformer
ACMeasurement	Can be used to represent electrical energies consumed by an electrical system which uses an Alternating Current (AC) for a three-phase (L1, L2, L3) or single-phase (L) and neutral (N). https://github.com/smart-data-models/dataModel.Energy/tree/master/ACMeasurement

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 3: Smart Data Models in transmission / distribution system.

Transmission / distribution system	
Smart Data Model	**Description and link**
Device	Allows to represent devices of different nature (IoT, mobile, wearable, etc.). Sensors/devices in the transmission/distribution system can be represented using this model. https://github.com/smart-data-models/dataModel.Device
DeviceModel	Captures the static properties of a Device. https://github.com/smart-data-models/dataModel.Device/blob/master/DeviceModel/README.md
Substation	Represents equipment through which electric energy in bulk is passed for the purposes of switching or modifying its characteristics. https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/Substation
PowerTransformer	Can be used to represent power transformers: electrical device consisting of two or more coupled windings, with or without a magnetic core, for introducing mutual coupling between electric circuits. https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/PowerTransformer
ACMeasurement	Can be used to represent electrical energies consumed by an electrical system which uses an Alternating Current (AC) for a three-phase (L1, L2, L3) or single-phase (L) and neutral (N). https://github.com/smart-data-models/dataModel.Energy/tree/master/ACMeasurement

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 4: Smart Data Models in energy consumption.

Energy consumption Smart Data Model	Description and link
Device	Allows to represent devices of different nature (IoT, mobile, wearable, etc.). Devices that consume energy can be represented using this model. https://github.com/smart-data-models/dataModel.Device
DeviceModel	Captures the static properties of a Device. https://github.com/smart-data-models/dataModel.Device/blob/master/DeviceModel/README.md
ACMeasurement	Can be used to represent electrical energies consumed by an electrical system which uses an Alternating Current (AC) for a three-phase (L1, L2, L3) or single-phase (L) and neutral (N). dataModel.Energy/ACMeasurement at master · smart-data-models/dataModel.Energy · GitHub
EnergyConsumer	Data model for users of energy or points of consumption on the power system model. https://github.com/smart-data-models/dataModel.EnergyCIM/tree/master/EnergyConsumer
EVChargingStation	Represents Charging Stations for Electric Vehicles. https://github.com/smart-data-models/dataModel.Transportation/tree/master/EVChargingStation
SmartMeteringObservation	Description of a Smart Meter Observation, generally applicable for Smart Homes, Industry, Cities and Agriculture. https://github.com/smart-data-models/dataModel.Device/tree/master/SmartMeteringObservation

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 5: Smart Data Models in environment and context.

Environment and context Smart Data Model	Description and link
AirQualityObserved	An observation of air quality conditions at a certain place and time. https://github.com/smart-data-models/dataModel.Environment/tree/master/AirQualityObserved
Building	Captures information on a given Building. https://github.com/smart-data-models/dataModel.Building/tree/master/Building
IndoorEnvironmentObserved	Observations of air and climate conditions for indoor environments. https://github.com/smart-data-models/dataModel.Environment/tree/master/IndoorEnvironmentObserved
WeatherForecast	Harmonised description of a Weather Forecast. https://github.com/smart-data-models/dataModel.Weather/tree/master/WeatherForecast
WeatherObserved	Observations of weather conditions at a certain place and time. https://github.com/smart-data-models/dataModel.Weather/tree/master/WeatherObserved

4.2 Existing vocabulary hubs

4.2.1 Vocol and VoCoReg

VoCol[21] is a project of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS - VoCol is an Integrated Environment for Collaborative Vocabulary Development in Distributed Version Control Systems.

Linked Data vocabularies are a crucial building block of the Semantic Data Web and semantic-aware data-value chains. Vocabularies reflect a consensus among experts in a certain application domain. They are thus implemented in collaboration of domain experts and knowledge engineers. Particularly the presence of domain experts with little technical background requires a low-threshold vocabulary engineering environment.

[21] vocol.iais.fraunhofer.de & www.vocoreg.com

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Inspired by agile software and content development methodologies, the VoCol methodology and tool environment addresses this requirement. VoCol is implemented without dependencies on complex software components, it provides collaborators with comprehensible feedback on syntax and semantics errors in a tight loop and gives access to a human-readable presentation of the vocabulary. The VoCol environment is employing loose coupling of different components for syntax validation, documentation generation, visualization, etc. on top of a standard Git repository. All VoCol components, even the repository engine, can be exchanged with little effort.

4.2.2 Semantic Treehouse

Semantic Treehouse[22] is an online community platform for semantic data models developed by TNO.

The platform combines the publication, maintenance, and governance for data models in one place. Semantic Treehouse is based on more than 10 years of experience with developing, maintaining, and sharing data standards. The platform can be branded and styled to a specific corporate identity for a recognizable user experience.

Different standardization bodies in the Netherlands already use the platform for their community management and maintenance of semantic specifications. TNO plans to open source the platform.

4.2.3 Open energy platform VP

The Open Energy Platform[23] (OEP) is a cross-projects initiative aiming to ensure quality, transparency, and reproducibility in energy system research. The initiative provides a collection of various tools and information that help working with energy related data focusing on Open Data and the provision of modelling results under open licenses. The OEP uses an extensive meta-data set based on the tabular data package specifications and the FAIR principles to improve the findability, accessibility, interoperability, and reuse of digital assets.

OEP has specified the Open Energy Meta-data[24] (OEMetadata) standard for energy meta-data that provides as extensive set of meta-data based on the tabular data package specifications and the FAIR principles. Attributes of the OEP meta-data model can considered in the BD4NRG meta-data description models. In more details, the OEP meta-data model contains multiple fields (keys) in a nested JSON structure and can be useful for establishing the BD4NRG meta-data descriptions. The OEP meta-data capture information on various aspects of a public dataset, including:

- descriptive information such as name, description, subject, keywords, publication date, URL, documentation, grant number in case the released data are associated with a research project, licenses, contributors, and

[22] www.semantic-treehouse.nl

[23] openenergy-platform.org

[24] github.com/OpenEnergyPlatform/oemetadata

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- details on the data itself including spatial and location information, time period covered in the data (temporal information), start/end points for timeseries data, format (e.g., CSV, XLS, JSON), units, fields, schema (structure of the data), decimal separator (symbol used to separate the integer part from the fractional part of a number written in decimal form), encoding.

A full list of attributes is available in the initiative's GitHub page[25]. Note that the OEP meta-data model can be customized, extended, and evolve to consider the needs of projects such as BD4NRG.

4.3 Conclusion

Given the large set of potentially useful data- and meta-data models, vocabulary hub functionality is essential for the BD4NRG ecosystem in achieving semantic interoperability. Information managers and software developers need to:

- have an overview of usable semantic models,
- know where to find and access those models, and
- have tooling available to start working with the models.

The semantic models are not only RDF ontologies, but also include XML schemas, taxonomies, code lists, mapping- (chapter 5) and validation specifications (chapters 5 and 6) and other documentation. All these specifications are published and maintained in different ways; some are published on GitHub, some have their own website, others are maintained and published by formal standardization bodies.

The Semantic Treehouse community platform as described in section 4.2.2. looks most promising to both providing overview and publishing links to the variety of existing standards as well as providing community support functionality to publish and maintain models that are governed by the BD4NRG ecosystem itself. Furthermore, the platform provides functionality to design data transformation and data validation specifications based on semantic models. These aspects of semantic interoperability are addressed in the subsequent sections.

[25] github.com/OpenEnergyPlatform/oemetadata/blob/develop/metadata/v150/metadata_key_description.md

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

5 Data transformation

The BD4NRG ecosystem is a wide data sharing agreement between many parties. It must therefore be able to handle and combine many different heterogenous data sets in different syntactical formats from or to different systems and APIs. Although shared semantic models allow every system to implement and speak the same language, this is not a realistic prerequisite in this case. In practices differences will continue to exist due to legacy implementations, different context/domains or, for historical reasons, competing standards. Furthermore, there is not always a positive business case for users to adapt a new semantic model in IT systems and interfaces.

A solution to overcome some of these barriers is to add data transformation services in the mix. However, the amount of required transformations in a network grows very quickly. That number is determined by the formula n x (n-1), where n is the number of nodes in a network. With many different models and versions, in combination with many linkages to other IT systems, this quickly becomes a complex and costly landscape to manage.

A better approach is to put a shared meta-data model in place. Rather than create translations (i.e., mappings) from each node to the other, each node is mapped centrally to this meta-data mode, thus reducing the number of translations required from n x (n-1) to 2n.

A key challenge that remains is to make the creation of translations (mappings) as easy and efficient as possible: from an ecosystems perspective, this requires separating the transformation logic from the logic that provides the data. This decoupling leads to translation software components, which are more easily reused by other users in the ecosystem. An example of a well-known method and tool that is traditionally used for this kind of software components is Extensible Stylesheet Language Transformations (XSLT).

Increased application of 'knowledge graphs', i.e., Linked Data, in the last decade[26] have given rise to another approach for data transformations: one based on formal semantics. Transformation components that are based on this approach enable users to make mappings between their data and some (hopefully shared) semantic model, i.e., an ontology. Doing this means the user has encoded the meaning of the data in question, thereby greatly increasing its reusability in the ecosystem. A method (incl. related tools) that has proven particularly useful for this purpose is the RDF Mapping language (RML).

5.1 Transforming semi-structured data into Linked Data

RML.io is a collection of methods (i.e., specifications) and tools (i.e. implementations) used to generate knowledge graphs from any type of semi-structured data. It is built by a team of researchers from Imec and Ghent University. There is support for CSV, JSON, and XML (formats) as well as multiple data sources, such as files, databases, Web APIs, and streams.

There are two main advantages of the RML suite. Firstly, it is open source and comes with adequate documentation. Secondly, it offers both specifications (method) as well as implementations (tools).

[26] https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

RML has been applied in numerous earlier H2020 projects such as INFINITECH[27], SPRINT[28] and ExtremeEarth[29].

5.1.1 RDF Mapping Language (RML)

At the core of the RML approach lies the writing of declarative rules using the RDF Mapping Language (RML). Through the execution of these rules, knowledge graphs are created from corresponding data sources using annotations provided through vocabulary terms. These vocabulary terms are derived from some ontology. The original data source remains unchanged.

Since it is declarative, RML rules are separated from the software that executes them, so the latter does not need to be updated when the rules are updated.

5.1.1.1 YARRRML and Matey

RML was built foremost with machine-processability in mind and, as the authors state[30], it is not always straightforward for users to define or understand these rules. New users, especially those already familiar with the YAML language, are advised to start with YARRRML[31], a human-friendly representation of RML. See also Matey[32], a browser-based YARRML editor, for a gentle introduction to creating knowledge graphs from using RML or YARRRML.

5.1.1.2 RML processors

RML rules need to be executed by some RML processor to generate a knowledge graph from (multiple) semi-structured data sources. As said, a strong point of the RML ecosystem is that it includes both specifications as well as implementations. Table 6 shows RML processors that are available at the time of writing[33].

[27] https://www.infinitech-h2020.eu

[28] https://sprint-h2020.eu

[29] http://earthanalytics.eu

[30] Heyvaert, Pieter, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. "Declarative Rules for Linked Data Generation at Your Fingertips!" In The Semantic Web: ESWC 2018 Satellite Events, 2018. https://doi.org/10.1007/978-3-319-98192-5_40.

[31] https://rml.io/yarrrml/

[32] https://rml.io/yarrrml/matey/

[33] https://rml.io/implementation-report/

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 6: Currently available RML processors.

Name	Version	Web page
RMLMapper	4.9.0	https://github.com/rmlio/rmlmapper-java
RMLStreamer	2.0.0	https://github.com/RMLio/RMLStreamer
CARML	0.3.0	https://github.com/carml/carml
RocketRML	1.0.6	https://github.com/semantifyit/RocketRML
SDM-RDFizer	3.2	https://github.com/SDM-TIB/SDM-RDFizer
Chimera	2.1	https://github.com/cefriel/chimera
Morph-KGC	1.4.0	https://github.com/oeg-upm/Morph-KGC

5.1.1.3 Limitations

The most pressing limitations to transforming data to Linked Data in general are: 1) mapping languages such as RML do not support complex data transformations, 2) the assumption that users are willing to learn and apply a dedicated mapping language, and 3) that users know of the original format (e.g., XML) as well as the target domain ontology before implementing any transformation.

The complex data transformations that are not covered by mapping languages are often performed with separate systems or through custom solutions.

As for the learning curve: the RML toolbox remedies the second limitation somewhat with YARRRML and its accompanying editor, Matey.

There are a few limitations to the RML approach specifically. Both have to do with the gap between the RML specification and actual implementations.

The first limitation is that “data velocity is not well supported in [RML] and corresponding processors, compared to data variety and volume”[34]. In other words, since implementations differ in how they handle varying data velocities, RML generated knowledge graphs are not always reproducible.

The second limitation is that the RML specifications “only partly align with Web APIs and streams descriptions”26, meaning the method doesn’t always cover Web API access and authentication, requiring developers themselves to add these extra steps.

[34] Van Assche, Dylan, Gerald Haesendonck, Gertjan De Mulder, Thomas Delva, Pieter Heyvaert, Ben De Meester, and Anastasia Dimou. “Leveraging Web of Things W3C Recommendations for Knowledge Graphs Generation.” In Web Engineering, 2021. https://doi.org/10.1007/978-3-030-74296-6_26.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

5.2 Conclusion

RML offers valuable methods and tools for facing the challenge of interoperability between BD4NRG use cases. It allows the BD4NRG pilots to develop knowledge graph construction pipelines that include the transformation of different types of content into RDF (i.e., Linked Data), without the need for custom parsers. Decoupling this Linked Data transformation capability is essential to drive data exchange and reuse in a wide data sharing agreement such as BD4NRG.

Two characteristics make RML particularly suitable: firstly, that it offers numerous building blocks, covering both their specification and implementation. Secondly, that it includes YARRML and Matey, making it approachable for developers without a background in Semantic Web.

The third limitation in the previous section points to an assumption in the BD4NRG reference architecture: that users have access to a meta-data catalogue that will tell them what data is available in what format, and that there is some vocabulary hub whose domain ontologies can be used to guide data transformations.

Hence, the BD4NRG toolbox will contribute to interoperability by enabling vocabulary hubs to publish transformation specifications with their vocabularies, and by making sure the meta-data catalogue provides sufficient information about data sources.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

6 Data validation

If data modelling is the creation of a shared language, data quality forms its spelling and grammar. For participants in a data space to efficiently work together and use each other's data it is important that they use the same data quality rules and have mechanisms in place to signal quality issues.

To a large extent, setting up data quality rules is done at the same time as creating a model, for example by specifying whether an entity attribute is mandatory, or what type of values are allowed. Depending on the syntax in which the data is expressed, different constraint languages are available for further specification of quality rules. For example, in the case of RDF data one can use SHACL (Shapes Constraint Language) for this purpose.

It is often difficult for organizations to suddenly implement a large set of strict data quality rules. Often legacy data is of low quality or newly generated data can still contain unexpected issues. A solution to these problems is a more gradual increase in data quality rules both in amount of constraints as well as in the severity of them.

Once data quality rules have been determined (and, if need be, a strategy for improving data quality is formed) the rules can be formalized and implemented/configured in software for automatic validation. There is plenty of open-source software available that allows developers to easily validate different syntaxes (i.e., formats) using related schema or ruleset specifications.

There are so many data formats that listing them all here isn't feasible. Nor would it be very useful, since many of them have uses that differ from our BD4NRG context, which focuses on data interoperability and exchange in multiple (sub)domains. The rest of this chapter is therefore dedicated to providing an overview for the most common data formats used for this purpose, and a partial list of tooling that is available for them (Table 7).

JSON, XML, CSV and RDF are four of the most used formats for data exchange. RDF is a little different because it is not bound by a single serialization format. RDF triples can be serialized with different syntaxes. Three of the most popular are RDF-XML, JSON-LD, and Turtle (likely the most human-readable popular serialization).

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 7: Subset of common media types[35]

Format	Extension	Mime Type
XML	xml	text/xml
JSON	json	application/json
CSV	csv	text/csv
RDF-XML	rdf	application/rdf+xml
JSON-LD	jsonld	application/ld+json
Turtle	ttl	text/turtle

6.1 Validation of common data exchange formats

JSON, XML, CSV and RDF are four of the most used formats for data exchange. RDF is a little different because it is not bound by a single serialization format. RDF triples can be serialized with different syntaxes. Three of the most popular are RDF-XML, JSON-LD, and Turtle (likely the most human-readable popular serialization).

XML and JSON are widely used to store and exchange data. Both come with a mature schema language for writing schema definitions: XML Schema Definition Language[36] (XSD) for XML, and JSON Schema[37] for JSON.

The purpose of schema definitions is to describe for some data exchange what fields are expected, and how the values are represented. XML schema's describe a type of XML document, JSON schema's describe a type of JSON document, and so on. The constraints they impose go beyond the basic syntactical constraints imposed by the data format itself.

CSV Schema[38] is a schema language for CSV. The UK National Archives introduced CSV Schema in 2014 as part of their efforts to preserve digital data with consistent meta-data. Although CSV Schema sees wider use, it is not as mature or widely adopted as XSD or JSON-schema. CSV Schema, like the CSV format itself, is relatively simple. While XSD is itself valid XML and JSON Schema's are proper JSON as well, a CSV Schema definition is not itself a CSV file but is text-based.

[35] www.iana.org/assignments/media-types
[36] https://www.w3.org/TR/xmlschema11-1/
[37] https://json-schema.org/
[38] https://github.com/digital-preservation/csv-schema

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

There are many software tools available for XSD and JSON Schema, including validators[39]. The creators of CSV Schema provide a validator as well[40]. Validators take some schema definition and one or more instances of data as input, test these instances against the provided schema, and generate a validation report as output. Reported errors are often in varying degrees of severity (e.g., either a 'warning' or an 'error').

Schematron deserves a special mention as it is one of the most powerful data validation tools available. It is a schema language for XML and often used in addition to XML Schema Definitions (XSD) for more powerful business rule validation. Rather than creating a grammar for an XML document like XSD, Schematron schemas make assertions about the presence or absence of patterns in XML trees. What Schematron can do is best illustrated by writing a typical Schematron in plain English:

"The Person element should in the XML instance document have an attribute Title and contain the elements Name and Gender in that order. If the value of the Title attribute is 'Mr' the value of the Gender element must be 'Male'."[41]

6.2 Validation of Linked Data

RDF is different from the before mentioned data formats and not just in that there are many serializations.

RDF's design is based on an open world assumption. As a result, validation has traditionally been difficult and less complete compared to the other data formats that we mentioned. Like XML and JSON, RDF has schema languages to describe the structure of RDF instance data. Unlike XSD and JSON Schema, RDF Schema (RDFS) is generally understood to supplement rather than validate RDF data. Instead, the standard for describing structural constraints on RDF data is the Shapes Constraint Language (SHACL).

RDF can be confusing for developers who have previously only seen JSON or XML structures, because the data has the structure of a graph, as opposed to the tree structures of JSON and XML. That means that the constraints written in SHACL are also in the shape of a graph. In fact, that is precisely what a SHACL 'schema' is called: a shapes graph.

However, the basic operation of SHACL processors is the same: they take inputs in the form of shapes graphs and data graphs. Both can be represented with RDF formats such as RDF/XML, JSON-LD or Turtle. The output of the validation process is a validation report with all the validation results. The validation report itself is also an RDF graph, with its own vocabulary[42].

[39] XML: https://www.w3.org/wiki/XML_Schema_software, JSON: http://json-schema.org/implementations.html

[40] https://github.com/digital-preservation/csv-validator

[41] Source: https://www.xml.com/pub/a/2003/11/12/schematron.html

[42] https://www.w3.org/TR/shacl/#validation-report

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

6.3 Examples

6.3.1 JSON Schema example

```
{
  "$id": "https://example.com/person.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "Person",
  "type": "object",
  "properties": {
    "firstName": {
      "type": "string",
      "description": "The person's first name."
    },
    "lastName": {
      "type": "string",
      "description": "The person's last name."
    },
    "age": {
      "description": "Age in years which must be equal to or greater than zero.",
      "type": "integer",
      "minimum": 0
    }
}
  }
```

Figure 10: JSON Schema definition example[43].

[43] https://json-schema.org/learn/miscellaneous-examples.html

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

6.3.2 XSD example

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

<!-- definition of simple elements -->
<xs:element name="orderperson" type="xs:string"/>
<xs:element name="name" type="xs:string"/>
<xs:element name="address" type="xs:string"/>
<xs:element name="city" type="xs:string"/>
<xs:element name="country" type="xs:string"/>
<xs:element name="title" type="xs:string"/>
<xs:element name="quantity" type="xs:positiveInteger"/>
<xs:element name="price" type="xs:decimal"/>

<!-- definition of attributes -->
<xs:attribute name="orderid" type="xs:string"/>

<!-- definition of complex elements -->
<xs:element name="shipto">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="name"/>
      <xs:element ref="address"/>
      <xs:element ref="city"/>
      <xs:element ref="country"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="item">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="title"/>
      <xs:element ref="quantity"/>
      <xs:element ref="price"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="shiporder">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="orderperson"/>
      <xs:element ref="shipto"/>
      <xs:element ref="item" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute ref="orderid" use="required"/>
  </xs:complexType>
</xs:element>
</xs:schema>
```

Figure 11: XML schema definition example[44].

[44] https://www.w3schools.com/xml/schema_example.asp

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

6.3.3 CSV Schema example

```
version 1.2
@totalColumns 3
name: notEmpty
age: range(0, 120)
gender: is("m") or is("f") or is("t") or is("n")
```

Figure 12: CSV Schema definition example[45].

6.3.4 Schematron example

```
<assert test="@Title">
The element Person must have a Title attribute.
</assert>
<assert test="count(*) = 2 and count(Name) = 1 and count(Gender)= 1">
The element Person should have the child elements Name and Gender.
</assert>
<assert test="*[1] = Name">
The element Name must appear before element Gender.
</assert>
<assert test="(@Title = 'Mr' and Gender = 'Male') or @Title != 'Mr'">
    If the Title is "Mr" then the gender of the person must be "Male".
</assert>
```

Figure 13: Example of a Schematron definition example[46].

6.3.5 SHACL example

```
ex:Alice
    a ex:Person ;
    ex:ssn "987-65-432A" .

ex:Bob
    a ex:Person ;
    ex:ssn "123-45-6789" ;
    ex:ssn "124-35-6789" .

ex:Calvin
    a ex:Person ;
    ex:birthDate "1971-07-07"^^xsd:date ;
    ex:worksFor ex:UntypedCompany .
```

Figure 14: SHACL shape graph example[47].

6.4 Conclusion

Validation of outgoing or incoming data is a basic building block for the BD4NRG platform. Without it, there is no way of verifying if data conforms to the expected structure and other requirements, which

[45] https://digital-preservation.github.io/csv-schema/csv-schema-1.2.html
[46] https://www.xml.com/pub/a/2003/11/12/schematron.html
[47] https://www.w3.org/TR/shacl/

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

will impede interoperability. Fortunately, there are many available open-source software components, i.e., building blocks, available on the web for this purpose.

The quickest way for BD4NRG pilots to start benefiting from available validation technology is by adopting existing vocabulary standards for their data in the first place. The reason being that most vocabulary maintainers also provide validation artifacts to drive the vocabulary's adoption by users. Adopting these standards in the BD4NRG pilots (provided they match the data sufficiently, of course) has the potential to improve at once both interoperability and data quality.

For the BD4NRG toolbox, this means two things: it must be able to publish and share data validation specifications in the vocabulary hub component. Secondly, it will need to provide a way for parties in the BD4NRG ecosystem to publish and search for data validation services in the catalogue.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

7 Conclusions

This BD4NRG project deliverable D4.2: 'Report on Existing Methods, Tools and Prototype Implementations' is the first report of BD4NRG Task 4.2. It has identified and assessed the various types of semantic tools as required in the subsequent phases of semantically managing the various artefacts in a federated and open data space approach as pursued by BD4NRG: meta-data brokering, providing vocabularies, data transformation and data validation.

7.1 Assessment of semantic building blocks for BD4NRG pilots

The current deliverable D4.2 has identified and assessed different types of semantic tools, or building blocks, that are required by BD4NRG. Four types of semantic building blocks were identified: semantic brokering, vocabulary providing, data transformation, and data validation. Table 8 summarizes the status of these building blocks, which of them are currently available for (re-) use in BD4NRG, and which of them should be developed within the BD4NRG project.

Table 8: Assessment of semantic building blocks for BD4NRG pilots.

Building blocks needed for BD4NRG Pilots	Assessment of availability for BD4NRG
Meta-data Broker	*What semantic building blocks are needed for BD4NRG pilots?:* Federated brokering for: • Data services available for AI applications • Machine learning models *What semantic building blocks are already available for BD4NRG (after small adjustments)?:* • IDS information model for describing analytics services. • DCAT-standard and DCAT-AP in particular can be used as a starting point for the conceptual model of the catalogue. *What follow-up will be done by for BD4NRG T4.2?:* • Extensions to describe analytics services being developed in BD4NRG across four themes (BD-4-NET, BD-4-ENEF, BD-4-DER)

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Building blocks needed for BD4NRG Pilots	Assessment of availability for BD4NRG
Vocabulary Hub	*What semantic building blocks are needed for BD4NRG pilots?:* For registration and management of: • Vocabularies (i.e., ontologies, reference data models, or meta-data elements) to annotate and describe data services available for AI applications, • Mappings between various vocabularies. *What semantic building blocks are already available for BD4NRG (after small adjustments)?:* • Smart data models based on the NGSI-LD information model. • Open Energy Family data description attributes. • Semantic Treehouse community platform. *What follow-up will be done by for BD4NRG T4.2?:* • Extensions of Smart data models to address the needs of BD4NRG pilots. • Extensions of Open Energy Family data attributes to describe the data assets of the BD4NRG pilots. • Further research the Semantic Treehouse community platform.
Data Transformation	*What semantic building blocks are needed for BD4NRG pilots?:* For handling many heterogenous data sets in different syntactical format: • Shared meta-data model. *What semantic building blocks are already available for BD4NRG (after small adjustments)?:* • RML. *What follow-up will be done by for BD4NRG T4.2?:* • Contribute to interoperability by enabling vocabulary hubs to publish transformation specifications with their vocabularies.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Building blocks needed for BD4NRG Pilots	Assessment of availability for BD4NRG
Data Validation	*What semantic building blocks are needed for BD4NRG pilots?:* • The ability to publish and share data validation specifications. • The ability for parties to search for data validation services in the catalogue. *What semantic building blocks are already available for BD4NRG (after small adjustments)?:* • JSON Schema • XSD • CSV Schema • Schematron • SHACL *What follow-up will be done by for BD4NRG T4.2?:* • Investigate and determine the applicability of the above mentioned validation methods in BD4NRG

7.2 Follow up

Based on the outcome of the assessment in Table 8, Task 4.2 in BD4NRG will develop the meta-data broker and the vocabulary hub. Both the architecture and implementation of the meta-data broker and the vocabulary hub will be realized and documented in the follow-up deliverables of BD4NRG Task 4.2, i.e.:

- D4.3 'Architecture document on the integration of selected existing methods, tools, platforms in the BD4NRG platform', to be delivered in M20,
- D4.5 'An implementation of semantic interoperable components in the BD4NRG platform for selected use cases in the energy domain', to be delivered in M27,
- and in Task 4.2's input for the deliverables of adjacent BD4NRG tasks, i.e.:
- D4.7, D4.8 'BD4NRG Processing / Analytics - Technology Releases', to be delivered in M21 and M28, respectively.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

References

[1] EU AI4EU project. "The EU AI on Demand Platorm". URL: https://www.ai4europe.eu/

[2] EU AI4EU project. "I-nergy – Energy Load Forecasting". URL: https://aiexp.ai4europe.eu/#/marketSolutions?solutionId=9fc0357c-2b50-4733-8225-44f78a9d5421&revisionId=ae6bd423-aa37-411f-a8f1-40aeb6b0bd4d&parentUrl=marketplace#md-model-detail-template

[3] EU AcumosAI project. "Making Artificial Intelligence Accessible To Everyone". https://www.acumos.org/

[4] EU EUHubs4Data project. "European Federation of Data Driven Innovation Hubs". URL: https://euhubs4data.eu/overview/

[5] EUHUBS4DATA Project Deliverable D4.4. "Service Interoperability Requirements". 2021.

[6] Bader, S., Pullmann, J., Mader, C., Tramp, S., Quix, C., Müller, A. W., Akyürek, H., Böckmann, M., Imbusch, B. T., Lipp, J., Geisler, S., & Lange, C. (2020). The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), The Semantic Web – ISWC 2020 (pp. 176–192). Springer International Publishing. URL: https://doi.org/10.1007/978-3-030-62466-8_12

[7] EU BD4NRG Project Deliverable D2.3. "BD4NRG Semantification & Interoperability". 2021.

[8] EU BD4NRG Project Deliverable D2.5. "BD4NRG Reference Architecture". 2021.

[9] Otto, B., Steinbuss, S., Teuscher, A., and Lohmann, S., IDSA (2019). "International Data Spaces: Reference Architecture Model Version 3," International Data Spaces Association – IDSA, URL: https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf

[10] EU OPEN DEI project. "Aligning Reference Architectures, Open Platforms and Large-Scale Pilots in Digitising European Industry". URL: https://www.opendei.eu/

[11] OPEN DEI. "Design Principles for Data Spaces – Position Paper". Version 1.0. April 2021, URL: https://design-principles-for-data-spaces.org/

[12] EU BIG IoT project Deliverable D2.4a. "High-level architecture specification". December 2016.

[13] Broring, A., Ziller, A., Charpenay, V., Thuluva, A. S., Anicic, D., Schmid, S., Zappa A., Linares M. P., Mikkelsen, L., Seidel, C. (2018) "Enabling IoT Ecosystems through Platform Interoperability". IEEE Pervasive Computing, 17/4, 2018, Page(s) 41-51, ISSN 1536-1268 2017. URL: https://ieeexplore.ieee.org/document/8628296

[14] EU INTER-IoT project Deliverable D5.1. "Design Patterns for Interoperable IoT Systems". December 2017.

[15] Ganzha, M., Paprzycki, M., Pawłowski, W., Szmeja, P., Wasielewska, K. (2017). "Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective". Journal of Network and Computer Applications, 81, 2017, Page(s) 111-124, ISSN 1084-8045. URL: https://www.sciencedirect.com/science/article/pii/S1084804516301618?via%3Dihub

[16] EU symbIoTe Project Deliverable D2.1 "Semantics for IoT and Cloud resources". 2015.

[17] Cimmino, A., et al. (2019). "VICINITY: IoT Semantic Interoperability based on the Web of Things" 15th International Conference on Distributed Computing in Sensor Systems (DCOSS) proceedings.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

[18] EU INTER-IoT project Deliverable D3.1 “Methods for Interoperability and Integration”. December 2016.

The BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.