D2.3: BD4NRG Semantification & Interoperability

WP2 – System Requirements and Specifications

July 2021

www.bd4nrg.eu

Disclaimer

The BD4NRG project is co-funded by the Horizon 2020 Programme of the European Union. This document reflects only authors' views. The European Commission is not liable for any use that may be done of the information contained therein.

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	www.bd4nrg.eu		
Project Coordinator	ENG		
Deliverable	D2.3 – BD4NRG Semantification & Interoperability		
Work Package	WP2 – System Requirements and Specifications		
Delivery Month (DoA)	M7	Version	1.0
Actual Delivery Date	30/07/2021		
Nature	Report	Dissemination Level	Public
Lead Beneficiary	TNO		
Authors	Efthimios Bothos [NTUA], Babis Magoutas [NTUA], Alberto Abella [FIWARE], Katharina Wehrmeister [RWTH], Stefan Dähling [RWTH], Leonardo Carreras Rodriguez [RWTH], Alexander Pastor [RWTH], Ali Buyuk [OEDAS], Ibrahim Gazioglu [OEDAS], Wouter van den Berg [TNO]		
Quality Reviewer(s):	Javier Valino [ATOS] Francesco Nucci [ENG]		
Keywords	Data models, interoperability specifications, metadata, data clusters		

Preface

BD4NRG focuses on addressing emerging challenges in big data management for 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 (EPES) 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 10 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
- De-risking investments in energy efficiency and increasing the efficiency and comfort of buildings BD-4-ENEF

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

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	
19	ASM TERNI SPA	**ASM**	IT	

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
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 DOO SISTEMSKI OPERATER PRENOSNEGA ELEKTROENERGETSKEGA OMREZJA	**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	

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

Contents

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

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

List of figures

List of tables

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

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

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

Executive Summary

Data interoperability is an important aspect of data sharing. This is especially true the more systems with varying interfaces are involved. Integrating non-interoperable systems is costly and scales poorly. As BD4NRG aims to handle very large amounts of data from diverse sources, this issue is crucial to the project.

The goal of this document is to provide an impression of the data landscape among BD4NRG's Use Cases, and to identify potentially useful existing efforts toward big data interoperability and semantification.

To this end, the IDS information model is regarded as a starting point. The portions of it which are related to semantics ("Context", "Content", "Concept") are introduced and examples presented in context of the three different clusters of BD4NRG LSPs.

The Deliverable goes on to discuss the importance of a metadata repository and the definition of a shared language respectively, relating the latter to the reality of the project's Use Cases. After covering some additional aspects of metadata, the corresponding chapter ends by mentioning the relevance of the topic to BD4NRG beyond just data resources, pertaining as well to the involved services, ML models and their respective interoperability. Specific analysis regarding this topic will be in scope for the later Deliverable 2.6.

To gain the overview of the data resources and data sets relevant within BD4NRG, this Deliverable presents the data used by each LSP (given the status of information available at time of writing) and attempts to find commonalities. For this purpose, the concept of "Data Clusters" is introduced. Classifying the relevant data sets in this way aids in making their diversity very clear, while also pointing out the overlaps that do exist.

The document closes by discussing the Agile Standardization Approach, followed by the introduction of some potentially useful standardizations and data models.

A direct connection between these potentially useful approaches and specific Use Cases is not made in this Deliverable. As the project is still in its early stages, and the data needs of LSPs are likely to be extended or modified as time goes on, such matchings are not possible at the time of submission. This topic will instead be addressed in BD4NRG Deliverable 2.6.

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. Purpose of this document

This document describes the results of the work done by the group of Task 2.4 – Semantic Data Models & Service and Platform Interoperability Specifications.

The words 'semantification' and 'interoperability' in the document title refer to the sharing of data. The more interoperable software programs are, the easier it is for them to exchange information. Software programs that are not interoperable need to be integrated before exchange of information can happen. Ad-hoc integration is often costly and, more importantly, doesn't scale well.

One of the goals of the BD4NRG Reference Architecture is to facilitate interoperability between data suppliers, data consumers and data service providers in BD4NRG. The goal of this document is to serve as a guide to semantic interoperability for both the designers of this architecture, as well as its prospected users.

1.2. Relation to other documents

First, this document is based on the content of D2.1, which served as a starting point for the analysis of the information usage and needs in the Large-Scale Pilots (LSPs) and their associated Use Cases (UCs).

Secondly, this document relates to D2.2 and D2.5 since these too describe aspects of the BD4NRG architecture, albeit from different perspectives. D2.2 is focused on the structure and functionality of the BD4NRG components (i.e., a 'bottom-up' approach) while D2.5 looks at the convergence of existing architecture efforts to form the different architectural layers of the BD4NRG platform (a 'top-down' approach).

Finally, there is a strong relation with Task 4.2, which is expected to implement the specifications that T2.4 provides in this document (D2.3) and those in D2.6. They will enable the discovery and interoperability of the various artefacts which reside in the B4NRG platform (not only Data, but Services, Edge Resources and ML models as well) for the stakeholders of the platform through semantic metadata definitions and API's.

1.3. Document structure

This document is structured as follows. Chapter 2 provides the basis of the BD4NRG semantic interoperability specifications. It starts with an outline of metadata categories based on the IDS Information Model. Since our purpose is to work towards semantification and interoperability, the focus is on the three metadata categories that have a strong link with semantics: Content, Context and Concept.

The last of those categories, Concept, points to the need for shared languages about the domain of interests in the BD4NRG platform. Chapter 3 presents an analysis of what these domains of interests are, based on the planned use of data in the current designs of the twelve LSPs. The result of this preliminary analysis is a clustering of this data into different domains, plus a look at the extent to which data usage overlaps between the LSPs.

This document concludes with chapter 4, which is a first introduction to the data model landscape. It describes a small subset of common data models that should fit one or more of the discovered data clusters. These models are therefore expected to be instrumental to achieve interoperability in the BD4NRG platform.

1.4. Looking forward

The current deliverable identifies the needs of data modelling in chapter 3, identifying precisely what are the main entities to be modelled. These entities are likely to be extended / modified when the actual implementation of the LSP will start. This document presents the grounds for starting the clustering.

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

On the other hand, in chapter 4 the semantic standards are presented as sources for a common modelling of the data of the LSP. Based on them, it will allow the adoption of those more convenient to match the global architecture defined in other tasks of the project and the modelling of the entities of the different LSP.

So, this deliverable is setting the grounds for a successful data interchange at Data spaces, because they will share a common access to data resources (common API and compatible identification providers) and a common semantic modelling of the data assets to be managed by the different LSP. Hopefully, the final solution will be based on open standards to allow a quick adoption and adaptation to other needs in the Energy Sector and beyond. These further insights about the solution will be covered in the next deliverable D2.6. Thus, Deliverable D2.6 will do this matching between these standards semantic sources mentioned in chapter 4 and the actual needs of the LSP allowing a de facto standardization.

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

2 Towards defining interoperability specifications

2.1. Introduction

2.1.1. Data space design principles

In a recent position paper, Open DEI described ‘data spaces’ as follows (Ahle et al., 2021)[1]:

“From a technical perspective, a data space can be seen as a data integration concept which does not require common database schemas and physical data integration, but is rather based on distributed data stores and integration on an “as needed” basis on a semantic level. Abstracted from this technical definition, a data space can be defined as a federated data ecosystem within a certain application domain and based on shared policies and rules. The users of such data spaces are enabled to access data in a secure, transparent, trusted, easy and unified fashion. These access and usage right can only be granted by those persons or organisations who are entitled to dispose of the data.”

The BD4NRG scope has been described in similar terms Indeed, the BD4NRG architecture has been described as open and highly distributed since the project’s inception a few years ago. Data space design principles should therefore be followed as much as possible and where appropriate.

2.1.2. Interoperability

Interoperability is the ability of two or more ICT assets (hardware devices, communications devices, or software components) to work together easily or automatically and, in the business sense, expands to include the ability of two or more business processes, or services, to easily or automatically work together.[2] Software programs that are not interoperable need to be integrated. Therefore, integration is a process that enables interoperability. One of the goals of the BD4NRG Reference Architecture, then, is to minimize the cost and effort required for integration of data stores for a particular purpose.

The quality of interoperability can thus be measured by integration effort. Standards help to increase the quality of interoperability by reducing the integration effort.[3] When systems are to be integrated to fulfil a function cooperatively, all interoperability concerns need to be covered. In many interoperability frameworks, these concerns are grouped in layers.

2.1.3. Interoperability in different layers

Interoperability frameworks such as SGAM and the EIF (European Interoperability Framework) apply a layered model to explain interoperability. At the organizational layers, interoperability means that organizations have overlapping objectives and processes, which form the basis of collaboration. At the technical layers, interoperability means using identical (or at least compatible) mechanisms for establishing connections and exchanging data through the use of ICT. In between these two layers is where ICT and business processes meet. In this ‘informational’ layer, data becomes information. Here, interoperability means that there’s a common understanding of the format and meaning of data between stakeholders so that it’s clear how it can be used in business processes, preventing ambiguities regarding the notions in the domain of interest. This is also referred

[1] Ahle, U., Bastiaansen, H., Bengtsson, K., Caballero, M., Castellvi, S., Dognini, A., Ette, F., Faraldi, M., Gelhaar, J., Graziani, A., Grguric, A., Gusmeroli, S., Helmholt, K., Hierro, J., Hoppenbrouwer, D., Hülsmann, T., Krco, S., & Valiño, J. (2021). *Design Principles for Data Spaces – Position Paper* (L. Nagel & D. Lycklama, Eds.). International Data Spaces Association.

[2] White Paper- European Interoperability Framework (2004)

[3] Smart Grid Reference Architecture (2012)

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

to as semantic interoperability. In data spaces, semantic interoperability requires data stores to publish semantic metadata.

2.2. Semantic metadata of data sources

The decentralized and distributed nature of data spaces have important consequences for the interoperability specifications. Data providers must be able to describe and publish their data sources in such a way that these become discoverable for (potential) data consumers, whoever these may be. Consequentially, the core of any data space interoperability specification must be an unambiguous metadata model that is understood and agreed upon by every participant in the data space (Bader et al., 2020)[4].

The IDS Information Model (see Figure 1) is an example of such a metadata model for data spaces. Since it is designed to cover the whole range of possible interactions in IDS, it goes beyond semantics. But, as we will see, it serves as a good starting point to explore requirements of semantic metadata in BD4NRG.

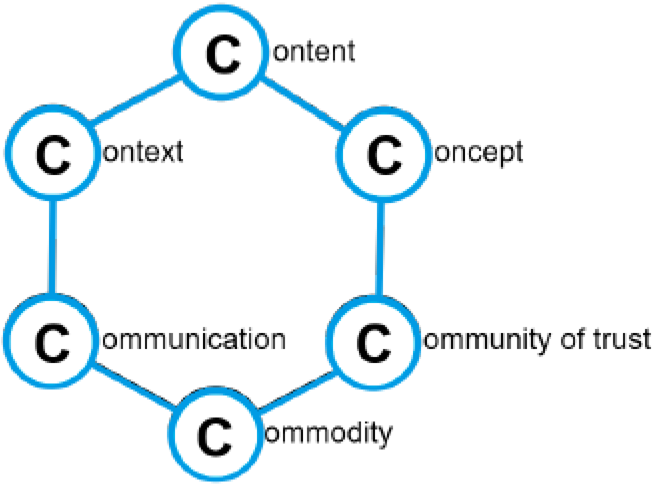

Figure 1 - basic visualization of the IDS Information Model

According to this model, there are three types of metadata that together describe the semantics of a digital data resource, i.e., 'what' the data is about: Content, Concept and Context (Otto et al., 2019)[5]:

"As a Resource's content is its most essential aspect, Content *is located at the top of the hexagon. This content is interpretable by references to a shared, formally defined* Concept*, whereas links to a particular* Context *(in terms of time, place, or real-world entities) make the content potentially relevant for certain Data Consumer."*

The three remaining concerns are: *Communication*, *Commodity* and *Community of Trust*. While interesting, they have less to do with semantics and more with other aspects of the BD4NRG architecture. Those aspects are addressed in other deliverables, notably D2.2, D2.5 and later in D2.6. Therefore, they remain out of scope here.

[4] 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. https://doi.org/10.1007/978-3-030-62466-8_12

[5] Otto, B., Teuscher, A., & Lohmann, S. (2019). *IDSA Reference Architecture Model*. International Data Spaces Association. https://www.opendei.eu/wp-content/uploads/2020/01/IDS-Reference-Architecture-Model-3.0.pdf

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

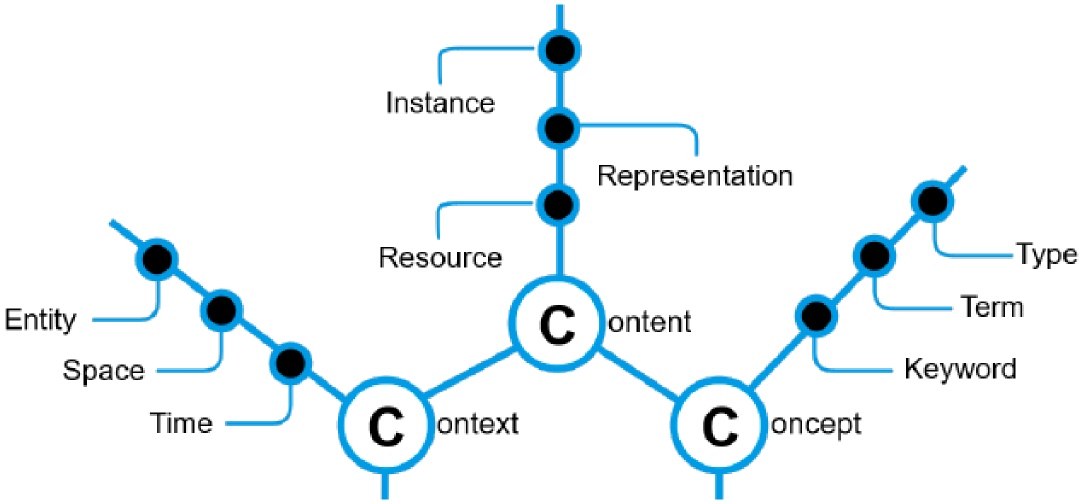

Figure 2 - A closer look at Content, Concept and Context

2.2.1. Content

The most basic elements of metadata are those that we use daily, e.g., file size, file format, file creation, etc. These concerns have to do with the Content of the data resource. The IDS Reference Architecture Model (RAM) describes Content as follows:

The Content concern deals with the description of a Resource's inherent substance, i.e., its "content" available in any machine- interpretable, binary format. It addresses questions like:

- *What type of content does a Resource provide (e.g., text or an image)?*
- *What does the content look like (e.g., what is its structure, format etc.)?*
- *Is a content sample provided?*
- *What is the size and creation date of a particular file?*

While Content metadata is not directly concerned with semantics, it is the starting point from which other metadata can be added.

2.2.2. Context

Context metadata is an important way for potential consumers of the data to assess its relevance to their needs. As per the IDS RAM:

The Context concern deals with temporal and spatial aspects as well as with real-world entities a Resource's content relates to (intrinsic context). It addresses questions like:

- *What time period does the content cover?*
- *When and where was it gathered?*
- *Which sub-entity of a larger entity does a certain dataset relate to?*

Context management allows the gathering of data from heterogeneous systems, and the setting of thresholds for alerts about the values of the data retrieved. It also allows, with the right configuration the federation of elements managing the context. This approach allows the system of systems approach and therefore the management of digital twins of the Energy systems. A well-known example that operationalizes this idea is the NGSI-LD standard, which is created under the European Telecommunications Standards Institute (ETSI), an international standardization body for context management.

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

2.2.3. Concept

Concept metadata is what helps computers 'interpret' the data, to tell what it really means. This is the core of semantic metadata. Adding concept-metadata opens a whole world of new possibilities for automation. There is a field within the area of artificial intelligence that is known as 'knowledge representation and reasoning' where technologies are developed that use semantic metadata to solve a host of complex tasks. The IDS RAM describes it as follows:

The Concept concern deals with the modelling of the "meaning", annotation, and interpretation of entities introduced by the orthogonal Resource concerns (Content, Context, Communication etc.). It addresses questions like:

- *What type of observation does the data refer to "temperature"?*
- *What kind of object does a context entity represent (factory, building)?*
- *What is the meaning of a certain date parameter (beginning or end of a range)?*

Keywords express the "meaning" of an entity via informal natural language tags. As keywords can be chosen freely by a Data Provider, they are prone to inconsistencies and errors. Using controlled vocabularies, it is possible to add curated, (formally) defined and reusable Terms, which can be shared across different scenarios and domains. In addition, conceptual schemas and ontologies define Types of entities, if these are to be individually modelled as custom instances.

2.3. Examples

What if data owners in the BD4NRG project would provide metadata about their data stores in the way described above? To see what that could look like, we've selected one example from each of the three BD4NRG LSP clusters and specified their metadata in a general, non-technical way. Note that this is for illustrative purpose only; the metadata below is completely hypothetical and only slightly based on the original LSP descriptions. The actual BD4NRG LSPs are visited in more detail in chapter 3.

2.3.1. BD-4-NET: example inspired by LSP 4

BD-4-NET use cases show how big data can be used to Increase the efficiency and reliability of the electricity network. The example below does not represent the LSP in question and is only meant to illustrate the use of metadata for a data source.

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

Content	• **Resource:** The resource constitutes time-series (i.e., records) of daily energy consumption of consumers behind a single transformer (i.e., households, industry, illuminations etc.) in test area. The data is aggerated to protect consumer privacy.
	• **Representation:** The data will be provided over an API system and will use the JSON format, containing the following data
	  ○ Transformer ID
	  ○ Geo-coordinates
	    ▪ medium voltage assets coordinates (e.g., transformator coordinates)
	    ▪ Region coordinates
	  ○ Power Factor (Cosθ)
	  ○ Time series:
	    ▪ Energy data (V, A, kWh)
	• **Instance**: this JSON instance has 200 objects, date of creation: 2021-01-01, time period covered: one day on 2021-06-30. file size: 50KB.

Context	• **Time:** The time period will vary between 3, 6 and 12 months depending on the resource. The sampling frequency varies between 5 minutes, 15 minutes, and 60 minutes. The default is 15 minutes. • **Space:** all meters are installed at projects in the region of the DSO zone from Eskisehir, Turkey. This data instance is an aggregate and so not from a single project or neighbourhood. • **Entity:** data describes different assets in the DSO region of Eskisehir.

Concept	• **Keys:** e.g. “power demand” , “power factor”, “transformer”, “consumption”
	• **Terms:** e.g.
	  ○ Power factor: https://dbpedia.org/resource/Power_factor
	  ○ Transformer: https://dbpedia.org/resource/Transformer
	• **Type:** instead of keywords or (unrelated) terms, every data property could be mapped to a data model or ontology (e.g. SAREF), thereby formalizing the semantics (i.e. meaning and relations) into a network of classes and properties.

Table 1: hypothetical metadata for a data source mentioned in LSP 4

As this first example shows, the metadata provides important information on what the data is about, even though our example is still rudimentary.

What this example doesn’t show yet is that mapping key concepts to some vocabulary can be done for the metadata itself, too. For example, the meaning of ‘Eskisehir’ could be made explicit by providing a URI , e.g. `https://dbpedia.org/resource/Eski%C5%9Fehir_Province`. Now we know that the data providers mean the whole province of Eskisehir, not just the city.

2.3.2. BD-4-DER: example inspired by LSP 5

BD-4-DER use cases aim at using big data for optimizing the management of assets (Distributed Energy Resources – DER) connected to the grid. The example below does not represent the LSP in question and is only meant to illustrate the use of metadata for a data source.

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

Content	• **Resource**: Data that this resource provides are datasets of time ordered sensor readings (i.e. records) of residential demand response assets • **Representation**: Data of this resource is tabular and is serialized as a comma-separated values (csv) file format using UTF-8-character encoding. No schema is provided. • **Instance**: this instance has 442 rows, date of creation: 2021-07-01, time period covered: 2021-06-01 up to and including 2021-06-30. file size: 84KB.
Context	• **Time:** the period covered by this resource spans from 2019-01-01 up to the present (continuously updated). There is no fixed sampling frequency as readings depend on Demand Response events. • **Space:** all residential assets are in `https://dbpedia.org/resource/Belgium` • **Entity:** readings are derived from various types of residential demand response assets, e.g. electrical vehicles, electric boilers.
Concept	• **Keys:** e.g. **“**demand response”, “flexibility”, ‘residential assets”, “state of charge”, “power setpoint” • **Terms:** e.g.: o Set point: `https://www.wikidata.org/wiki/Q1752097` o State of charge: `https://www.wikidata.org/wiki/Q5368536` o electrical vehicle: `https://www.wikidata.org/wiki/Q1076882` o electric boiler: `https://www.wikidata.org/wiki/Q5357557` **Type:** instead of unrelated terms, every data property (i.e. column) is mapped to an ontology (e.g. SAREF), thereby formalizing the semantics (i.e. meaning and relations) of these properties.

Table 2: hypothetical metadata for a data source mentioned in LSP 5 (use case 1)

2.3.3. BD-4-ENEF: example inspired by LSP 12

BD-4-ENEF use cases describe big data driven solutions for de-risking investments in energy efficiency and increasing the efficiency and comfort of buildings. The example below does not represent LSP 12 and is only meant as another illustration of the use of metadata for a data source.

Content	• **Resource**: This resource constitutes time-series (i.e. records) of daily energy consumption for house heating. • **Representation**: Data of this resource is tabular and is serialized as either an XLSX file format or a comma-separated values (csv) format using UTF-8-character encoding. Each record (row) consists of four fields. The columns are labelled: ‘datetime’, ‘average heat energy consumption (kWh)’, ‘max’, ‘min’ • **Instance:** this instance has 13.032 records, date of creation: 2021-07-01, time period covered: 2021-01-01 up to, but not including, 2021-07-01. file size: 565 KB.
Context	• **Time:** This resource covers a period from 2021-01-01 till present and will continuously be updated until at least 2023-01-01 • **Space:** all meters are installed at projects in the region of `dbr:Gulbene_Municipality` This data instance is from a single project at 57°10'20.6"N 26°45'36.4"E

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

Content	• **Resource**: This resource constitutes time-series (i.e. records) of daily energy consumption for house heating. • **Representation**: Data of this resource is tabular and is serialized as either an XLSX file format or a comma-separated values (csv) format using UTF-8-character encoding. Each record (row) consists of four fields. The columns are labelled: 'datetime', 'average heat energy consumption (kWh)', 'max', 'min' • **Instance:** this instance has 13.032 records, date of creation: 2021-07-01, time period covered: 2021-01-01 up to, but not including, 2021-07-01. file size: 565 KB.
	• **Entity:** data describes properties of the building at Ābeļu iela 2, Gulbene, Gulbenes novads (Cadastre number: 50010010111)
Concept	• **Keys:** e.g. "heating", "energy consumption", "projects", "sensor readings" • **Terms:** e.g. o `Project`: https://dbpedia.org/ontology/Building • **Type:** e.g.: using RML[6], the data provider mapped these sensor readings to the concept `saref:Measurement`, which relates to property `saref:Power` and is measured in kWh (a `saref: PowerUnit`). The measurement is made by a `saref:Device`, that `saref:hasManufacturer` who is known as Metbox.

Table 3: hypothetical metadata for a data source mentioned in LSP 10

As you can see, the content of the data in the above example has perhaps a similar structure to the previous examples since they both consist of sensor readings, but the Context and Concept metadata is different. Also, in this example we included a first mapping of the data to the SAREF ontology under 'Type'.

2.3.4. An operational example

As we look at these three previous examples, we might start to notice that useful metadata is missing, such as data about the publishing party or the endpoint where the data is served. Indeed, these pieces of information are useful and should be included in the end, however since they are examples of metadata categories of the lower half of Figure 1 we have left these out for now.

Next to other metadata, it's also likely that we start to wonder what this metadata looks like in practice (meaning: in code).

Below is an IDS example copied from Bader et al. (2020) that is meant to show a more technical view of metadata required to make a data resource discoverable. The example concerns stock market data. The metadata correspond to the categories seen in the other examples, with some additions such as `publisher` and `resourceEndpoint`. The `ids`-prefix refers to the underlying IDS Information Model, but this is not required. The syntax of this example is Turtle, which is very common in the world of semantic technology.

[6] RML stands for *RDF Mapping Language*, see: https://rml.io/specs/rml/

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

```
_:StockData a ids:DataResource ;
ids:title "Wall Street Stock Prices 2019"@en ;
ids:description "This dataset contains the complete stock market prices
of all 2019 Wall Street listed companies by milliseconds."@en ;
ids:keyword "stock price", "Wall Street", "2019" ;
ids:publisher <http://idsa.org/participants/BusIntInc>;
ids:temporalCoverage [ a ids:Interval ;
ids:begin [ a ids:Instant ;
ids:dateTime "2019-01-01T00:00:00.000-04:00"^^xsd:dateTimeStamp ];
ids:end [ a ids:Instant ;
ids:dateTime "2019-12-31T23:59:59.999-04:00"^^xsd:dateTimeStamp]];
ids:language idsc:EN ;
ids:representation [ ids:instance _:StockDataCSV ; ids:mediaType
<https://www.iana.org/assignments/media-types/text/csv>];
ids:resourceEndpoint [ a ids:ConnectorEndpoint ; ids:accessURL
"https://bi.com/connector/reports/2019_wall_street.csv"^^xsd:anyURI];
ids:contractOffer _:StockDataOffer .
```

Figure 3: stock market data modelled as an IDS Data Resource (Bader et al., 2020)

By providing metadata of a data resource in *some* structured way (of which the above is a recent example), we create the necessary conditions for data discoverability. However, the metadata alone is not sufficient. Another necessary condition is that potential data consumers know where to look for these data resources. This brings us to the necessity of a metadata repository.

2.4. Metadata repository

In order for participants of a data space to use or access each other's data or deployed data services one has to be able to know of their existence and know where to find them. Thus, some searchable metadata has to be published such as described in the previous sections. Such a metadata repository of the data space allows a participant to search somewhere for the data sources accessible on the network. The richer the metadata, the more powerful the search capabilities of this repository. Data providers provide metadata to enable searches that target a specific metadata category, for example based on the location of the data or which data model is used.

It's possible to organize the metadata repository as a federated system. In a federated repository there is no central integration of metadata. Using a distributed query mechanism, a federated repository is functionally indistinguishable from a single repository for the user. It acts as a virtual catalogue that is overarching the whole network. A federated repository decomposes a search query posed by a data consumer into subqueries and sends them to the individual repositories systems (sometimes called brokers). After processing the search query, the system composites the results back into one result and sends it back to the data consumer.

Being able to formally express the semantics of both the available data as well as the search queries is a critical enabler of federated metadata repositories.

2.5. The use of shared language

Conceptual schema, ontology, data model, vocabulary, these are all means to achieve a shared language between stakeholders. Semantic interoperability can be defined as the existence and adoption of a shared language between system components/participants about *a domain of interest*.

Chapter 3 shows a large variety of data clusters in the BD4NRG LSPs. Although the focus is on Energy, there are still many different domains of interest are involved. It is a clear that a multitude of shared languages are expected and indeed required.

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

Developing a shared language is in part a technical exercise considering machine readable data formats, flexible data models and existing modelling standards. But, in a large part coming to a shared understanding is also an exercise in cooperation between stakeholders who together have to define clear and unambiguous concepts and relations which describe their domain. This domain model can then later be mapped onto another domain model to enable translation and interaction between the two domains. Once a model is agreed upon the shared language can be formally modelled and serialized using for example RDFS/OWL. We come back to the issue of developing shared languages in chapter 4.

It is unrealistic to expect users of the BD4NRG data space to all adopt a single common semantic data model, even if their domain of interest overlap. Therefore, one of the key mechanisms that the informational layer in the BD4NRG architecture should support is semantic conversion. This can be envisioned as *a type of service* that need to be present in the BD4NRG data space. Such a service may be developed for specific semantic conversions or for enabling easy-to-use mapping between semantic models.

Semantic conversion can also be understood as the act of mapping the original data source to a certain semantic model in order to provide (enriched) metadata. In that case this step is part of the onboarding process that connects the data source to the data space.

When onboarding a data space as a data provider, it has to be clear which shared languages (i.e., vocabularies) are in use before configuring the metadata of your data sources. This can be organized by the establishment of an intermediary role in the data space, namely that of a Vocabulary Provider. The Vocabulary Provider offers shared languages such as ontologies, conceptual schemas, or other semantic standards for use in the data space. By positioning themselves as a Vocabulary Provider, owners and maintainers of these semantic standards declare their commitment to the adoption of their language in the data space.

2.6. Other metadata

The following metadata categories have their use in the metadata repository and are in some ways related to the issue of semantics.

2.6.1. Data provenance

One risk when organizations want to engage in a deeper cooperation by sharing data with each other in an ecosystem is that they don’t know where the other party's data is coming from. This can lead to lower trust in data and in less value being created from it. Moreover, when multiple parties have data stating contradicting facts one has to know where this difference comes from and which statement can be used. E.g., when multiple parties in a data space provide a list of countries in the world, some may mention a country with disputed sovereignty status while others do not. As such, data provenance is an important factor to consider when engaging in data sharing. Depending on the type of data space being build provenance can be covered in different ways. One way is by strictly vetting participants and making sure that the data they provide cannot be contradictory to data already present in the ecosystem. In that case provenance is guarded on an organizational and/or legal level and members of the data space know they can trust data from other members since they passed the initial vetting procedure. A data ecosystem working in this way is a rather closed ecosystem of which every participant is carefully considered before being allowed to participate. This is an example of how choices in data space governance influence choices in the technical architecture.

Alternatively, a more open ecosystem is possible. In this case provenance can be covered on the data semantics level. This means that all data needs associated meta data describing where it comes from and who is responsible for it. This then allows resolving conflicts as they occur and considering the data application in the conflict resolution. This approach reduces the reliance of a central vetting or certification party.

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

2.6.2. Data quality

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 systems are set up to cleanse data which does not (yet) adhere to the data quality rules. Setting up data quality rules can be done at the same time as creating a model, for example by asking whether an attribute being included in the model is mandatory, or what type of values are allowed. 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 this is the gradual increase in data quality rules both in number of constraints as well as in the severity of them. Additionally, there are solutions to automatically improve data quality to some extent, using tools such as the LOD-Laundromat. 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 modelled. One way of doing this is by including them into your data model. A data model based on semantic web technologies such as RDFS[7]/OWL[8] can use SHACL[9] as a constraint language for example.

2.7. Beyond data sets: data services, edge-resources, ML models

This document is focused entirely on data resources. But this is just one half of the BD4NRG platform. The other half consists of services such as data services, edge-resources, and ML models. Metadata plays an essential role for interoperability of services, too. Not just about the data sources these services process, but also the metadata that describes the service contract, i.e., what it does.

If BD4NRG services process similar data resources but apply differing data models, then this will lead to more development effort, higher design complexity, and higher processing time at the service provider side. For service consumers it will mean they are hampered in achieving their goals because of unnecessary limitations in service interaction and composition. This underlines the importance of using standardized data models where possible.

The second point concerns the BD4NRG service contracts, i.e., a description of what the service can do. If services have similar capabilities but express these capabilities using very different schema's then this too hurts the interoperability within the BD4NRG platform. To combat this, BD4NRG participants should be encouraged to 1) separate the schema part of the service contracts for maximum reusability, and 2) harmonize these contracts by following naming conventions

Since so little was known about the BD4NRG services design at the time of writing, we decided to keep this subject out of scope for this document and focus on data resources. The interoperability specifications for BD4NRG services will be developed as part of the work towards D2.6, the final BD4NRG architecture specification.

2.8. Conclusion

This chapter lays the foundation for an open, distributed BD4NRG platform that supports all the metadata categories needed for users to achieve semantic interoperability as best they can. For this, shared languages (also called vocabularies or ontologies) are essential. Every domain of interest in the BD4NRG project will need a shared language. To get an overview of what these domains of interest are requires a closer look at the LSPs, including the data usage there. The next chapter provides such an overview.

[7] RDF Schema: `https://www.w3.org/TR/rdf-schema/`
[8] OWL, or Web Ontology Language: `https://www.w3.org/TR/owl-features/`
[9] SHACL, or Shapes Constraint Language: `https://www.w3.org/TR/shacl/`

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

3 BD4NRG Data clusters

This chapter focuses on understanding the data available from the BD4NRG large scale pilots, aiming to provide a deeper understanding of the required data models for a common representation across pilots. A data template was shared among pilot partners, which was used to register available datasets relevant for the implementation of the specific use cases and that will be used within the project. This process led to a record of a number of datasets which were analysed for commonalities in terms of focus and attributes. The template that was used is reported in Annex 5.3, whereas the analysis of the datasets is provided in the following sections.

3.1. BD4NRG Large Scale Pilots (LSPs): Overview of data availability

The overall BD4NRG service analytics reference framework will be demonstrated in 12 Large Scale Pilots, implemented in 10 countries. Rationale of pilot applications is to address in a combined way two of the major challenges actually hindering big data analytics value capturing smart energy grids:

1. Nurturing the shifting towards predictive / prescriptive analytics.
2. Enabling multiple data sources analytics for multiple applications.

Figure 1 below shows the geographical location of the pilots and the type of application to be demonstrated.

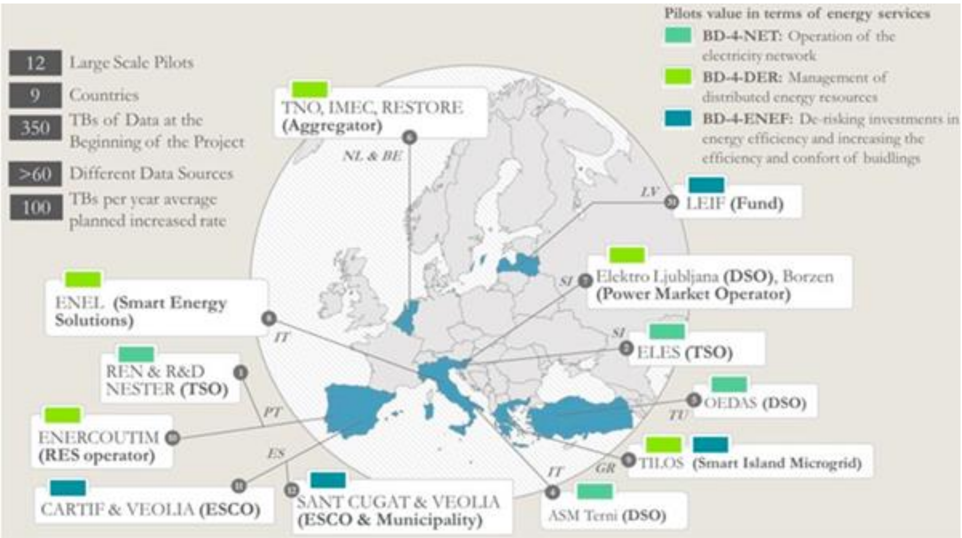

Figure 4: Overview of the 12 BD4NRG LSPs location.

The BD4NRG pilots aim to analyse different components of the energy system, in particular 3 types or clusters of LSPs are distinguished as follows:

- BD-4-NET LSPs will showcase how big data can be used to increase the efficiency and reliability of the electricity network;
- BD-4-DER LSPs aim at using big data for optimizing the management of assets (Distributed Energy Resources – DER) connected to the grid;
- BD-4-ENEF LSPs will provide big data driven solutions for de-risking investments in energy efficiency and increasing the efficiency and comfort of buildings.

The following sections provide an overview of that data that will be used in BD4NRG use cases. Note that as the project implementation progresses more data could be added and used where needed.

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

3.1.1. LSP 1: Integrating off-grid data with Condition based Monitoring for Enhanced Asset Maintenance

LSP 1 is a BD-4-NET LSP. It concerns the difficulties in Transition System Operator (TSO) data analytics that arise from the recent increase in data streams and sources. In order to derive predictive models for asset maintenance, improved data analytics tools are necessary.

The two Use Cases (UCs) in this LSP, coordinated by partners REN and R&D Nester, aim to assess the condition of Circuit Breakers (CB) and Overhead Lines (OHL) respectively. They look to provide answers to the following challenges in asset management:

- Implementation of optimal condition-based maintenance plans
- Incident characterization
- Predictive maintenance and implementation of maintenance plans

The following data sets will be considered:

1. **Circuit Breaker Incident Data:** This data is fetched via https once per CB trigger incident and stored in ASCII format. The readings are composed of the following attributes:

Attribute	Unit
Voltage Analog Channels	V
Current Analog Channels	A

Table 4 - LSP 1 data set: circuit breaker incident data

2. **Transmission Tower LiDAR Scan Data:** This data is fetched via https once per LiDAR scan and stored in LAS format. The readings are composed of the following attributes:

Attribute	Unit
3D Point Coordinate	m

Table 5 - LSP 1 data set: transmission tower LiDAR scan data

3. **Overhead Lines Right-of-Way LiDAR Scan Data:** This data is fetched via https once per LiDAR scan and stored in LAS format. The readings are composed of the following attributes:

Attribute	Unit
3D Point Coordinates	m

Table 6 - LSP 1 data set: Overhead lines right-of-way LiDAR scan data

3.1.2. LSP 2: Cross-functional Integration of Grid Maintenance Planning with Predictive Asset Management

LSP 2 is a BD-4-NET LSP, which focuses on the increased complexity of grid management and outage planning caused by increasing stochastic power generation behaviour with high penetration of renewable energy sources.

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

The primary goal of the related UC is to develop a decision support tool to help the experienced human planners in maintenance scheduling process. Consequently, the outage planner will need less workhours to produce an outage plan.

This decision support tool will help TSO in its ambition towards fewer outages, reduced outage duration, better outage utilization for maintenance purpose, increased system reliability, easier and faster outage rescheduling, better coordination with outside partners, possibility of dealing with more complex outages and therefore operation of the power system closer to the operational limits.

The BD4NRG partner responsible for this pilot is ELES.

The following data sets will be considered:

1. **Technical data:** This data is fetched via https once per purchase and stored in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
Manufacture date	Year
Asset Type	/
Type of cooling	/
Rated apparent power	MVA
Rated voltage for transformers	kV
Maximum operating voltage	kV
Rated Current	A
Manufacturer	/
Rated voltage for transmission line	kV

Table 7 - LSP 2 data set: technical data

2. **SCADA data:** This data is fetched via https & archive. The frequency can amount to 1 Minute, 15-minute, 1 Hour, or 1 Day. As an exception, the power transformer tap changer operation is sampled once per event. Data is stored as CSV Zip. The readings are composed of the following attributes:

Attribute	Unit
Average phase to phase voltage	kVrms
Average phase current	Arms
Average 3-phase active power	mW
Average 3-phase reactive power	MVAr

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

Attribute	Unit
Power transformer tap changer operation	Tap number

Table 8 - LSP 2 data set: SCADA data

3. **OLM data:** This data is fetched via https once every 1-15 minutes and stored in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
Apparent power	MVA
Active power	mW
Reactive Power	MVAr
Oil temperature	°C
Hotspot temperature	°C
Ambient temperature	°C
Phase voltages	kV
Tap changer position	/
Number of tap changer switching operations	/
Interphase voltage	kV
Operating frequency	Hz
Core temperature	°C
Dissolved Carbon Monoxide (CO) content	ppm
Dissolved Hydrogen (H2) content	ppm
Tap changer oil temperature	°C

Table 9 - LSP 2 data set: OLM data

4. **DGA measurements:** This data is fetched via https once per year and stored in an SQL database. The readings are composed of the following attributes:

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

Attribute	Unit
Dissolved Acetylene (C_2H_2) content	ppm
Dissolved Ethylene (C_2H_4) content	ppm
Dissolved Ethane (C_2H_6) content	ppm
Dissolved Methane (CH_4) content	ppm
Dissolved Hydrogen (H_2) content	ppm
Dissolved Carbon Monoxide (CO) content	ppm
Dissolved Carbon Dioxide (CO_2) content	ppm
Dissolved gasses together	ppm

Table 10 - LSP 2 data set: DGA measurements

5. **Oil quality measurements:** This data is fetched via https once every 3 years and stored in an SQL database. It serves to monitor the estimated lifetime of each transformer. The readings are composed of the following attributes:

Attribute	Unit
Colour grade	/
Oil flash point	°C
Refractive index	/
Acid number	mg KOH/g
Water content	ppm
Dielectric strength	KV 2,5mm gap
Interfacial surface tension	mN/m
Inhibitor content	%
Furan (2FAL) content	ppm
Furan (5HMF) content	ppm

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

Attribute	Unit
Furan (2FOL) content	ppm
Furan (2ACF) content	ppm
Furan (5MEF) content	ppm

Table 11 - LSP 2 data set: oil quality measurements

6. **Power transformer electrical measurements:** This data is fetched via https once every 3 years and stored in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
DC insulation resistance	MOhm
Insulation polarization index	/
Loss factor	%
Capacitance	pF

Table 12 - LSP 2 data set: power transformer electrical measurements

7. **Work orders:** This data is fetched via https once per revision or maintenance occurrence and stored in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
Maintenance Data	/

Table 13 - LSP 2 data set: work orders

3.1.3. LSP 3: Cross-Domain Cross-functional Integration of Predictive Asset Management for Grid Operation and coordination with flexible EVs charging

LSP 3 is a BD-4-NET LSP aiming to address the challenges Distribution Grid Operators (DSOs) face regarding asset maintenance due to the increased penetration of renewable energy sources and electric mobility.

On the one hand, asset maintenance is deeply linked to BD analytics of the sensors deployed in the network; on the other hand, it could be enhanced by leveraging active flexibility sources in the context of DR programs. In both cases, the amount of data that a DSO will collect and treat, will be significantly higher than the current datasets managed and stored by utilities and their IT infrastructures. BD analytics will definitely boost the DSO as one of the main actors for supporting energy transition and customer-centric energy markets.

The UCs in this LSP have differing goals: one will define and implement a new plan for asset replacement and maintenance based on predictive analytics. The other will define and implement optimal scheduling and usage

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

of RES and EVSE for reducing equipment loading during normal service, supporting DSO during the transition phase if necessary.

The BD4NRG partner responsible for this pilot is ASM.

The following data sets will be considered:

1. **Weather data:** This data is fetched via https 18, 24, and 5 times a day respectively and stored as SQL, csv or json respectively in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
Temperature	°C
Humidity	%
Wind speed	m/s

Table 14- LSP 3 data set: weather data

2. **Charging station data:** This data is fetched via WebSocket once per second and stored in json format. The readings are composed of the following attributes:

Attribute	Unit
Charging station ID	/
Charging station electric current	A
Socket ID	/
Socket status	/
Charging session ID	/
Charging session Start Time	/
Charging session End Time	/
Charging session energy	KWh
Charging session cost	€

Table 15- LSP 3 data set: charging station data

3. **Electric Vehicle data:** This data is fetched via TCP/IP once every 5 seconds and stored in json format. The readings are composed of the following attributes:

Attribute	Unit
Electric Vehicle ID	/

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

Attribute	Unit
Electric Vehicle Model	/
Connector Type	/
Battery Capacity	kWh
Battery Power	kW
Timestamp	/
SoC	%
Latitude	GPS
Longitude	GPS
Speed	km/h
Kilometers, autonomy	km
Odometer	km

Table 16- LSP 3 data set: electric vehicle data

4. **Customer energy data:** This data is fetched via the MQTT protocol once every 5 seconds. The readings are composed of the following attributes:

Attribute	Unit
Active power	KW
Voltage	V
Current	A
Reactive power	kVAr

Table 17- LSP 3 data set: customer energy data

5. **Secondary substation data:** This data is fetched via the MQTT protocol once every 0,2 seconds. The readings are composed of the following attributes:

Attribute	Unit
Voltage	V
Current	A

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

Table 18- LSP 3 data set: secondary substation data

6. **Primary substation data:** This data is fetched via the MQTT protocol once every 0,2 seconds. The readings are composed of the following attributes:

Attribute	Unit
Voltage	V
Current	A

Table 19- LSP 3 data set: primary substation data

7. **Data from acoustic sensors:** These sensors will be installed within the duration of BD4NRG and will produce WAV files with acoustic information at the rooms where transformers are installed. The aim will be to analyse this kind of data in order to predict transformer failures.

3.1.4. LSP 4: Predictive maintenance and fault prediction for MV grid and cross functional predictive asset management

LSP 4 is a BD-4-NET LSP aiming to address the challenges Services for the power distribution network – DSO. The related use case will be demonstrated through the implementation of maintenance and fault prediction operations.

The pilot takes place in medium voltage network distribution network of OEDAS. Pilot locations will be set after the final internal evaluation, but selection of an old network section with lots of MV faults in urban district will be considered.

OEDAS DSO does not have a professional maintenance plan and data evaluation processes, some unexpected faults on MV grid happen. To reduce faults and have more accurate predictive maintenance processes, the BD4NRG AI tools will initially work on long term historical data to analyse the accuracy of its short-term historical data. This fine-tuning process will serve as a continuous feedback loop for better prediction performance. Once the AI achieves a certain level of accuracy then the algorithm will be used to predict and analyse short term future maintenance and failure of the MV network assets, and hence failure along the grid. According to plan created model will be trained on one asset (preferably for transformer or any other asset which has historical data) and then same pre-trained model is further trained on different assets.

The BD4NRG partner responsible for this pilot is OEDAŞ.

The following data sets will be considered:

1. **Weather Data**

Attribute	Unit
Temperature	°C
Humidity	%
Wind speed	m/s

Table 20 - LSP 4 data set: weather data

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

2. **SCADA Analogue:** This data is fetched via WebSocket once per second and store in json format. The readings are composed of the following attributes in form of a float value:

Attribute	Unit
Van Phase neutral	kV
Vbn Phase neutral	kV
Vcn Phase neutral	kV
Vab Phase Phase	kV
Vbc Phase Phase	kV
Vca Phase Phase	kV
P Active power	MW
Q Reactive power	MVAR
S Apparent Power	VAR
f frequency	Hz
PF Power Factor	kV

Table 21 - LSP 4 data set: SCADA analogue

3. **SCADA:** This data is fetched via WebSocket too, once per second and stored in json format. The readings are composed of the following attributes in form of an integer value:

Attribute	Unit
Van Phase neutral	kV
Vbn Phase neutral	kV
Vcn Phase neutral	kV
Vab Phase Phase	kV
Vbc Phase Phase	kV
Vca Phase Phase	kV
P Active power	MW
Q Reactive power	MVAR

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

Attribute	Unit
S Apparent power	VA
f Frequency	Hz
PF Power Factor	kV

Table 22 - LSP 4 data set: SCADA

For the Predictive maintenance and fault prediction for MV grid and cross functional predictive asset management it is planned to use following Data from SCADA Circuit Breakers (CBs). Note that these data refer to alarms that are provided by the SCADA system:

Alarm type
Relay Malfunction
Relay Communication Error
Relay General Tripping Information
Sudden Over Current-1 Trip Information
Sudden Over Current-2 Trip Information
Delayed Overcurrent Trip Information
Sudden Gound-1 Opening Information
Sudden Ground-2 Opening Information
Delayed Ground Opening Information
Directional Overcurrent Opening-1
Directional Overcurrent Trip-2
Directional Overcurrent Opening-1
Directional Ground Overcurrent Opening-2
Trip Circuit Malfunction (TCS)
Thermal (Winding Temperature) Led
Thermal (Winding Temperature) Alarm

Table 23 - LSP 4 data set: SCADA CBs

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

3.1.5. LSP 5: Cross-stakeholder Transfer Learning for Flexibility Assets Forecasting and aggregation

LSP 5 is a BD-4-DER LSP, dealing with residential assets such as electric vehicles (EV), electric boilers, electric storage heaters, batteries, micro-CHP's, or heat pumps. They pose an untapped potential of flexible energy to integrate renewable energy by providing balancing services and mitigate grid congestion.

The UC related to this LSP focuses on the automation and user experience necessary to address upcoming customer needs, and is split in two subtopics:

1. Transfer learning, transferring knowledge characterized by a control policy across similar assets in different geographical locations, device types or customer segments
2. Cross-context grid friendly DR.

The BD4NRG partner responsible for this pilot is Centrica Business Solutions Belgium.

The following data sets will be considered:

1. **Residential demand response assets:** This data is fetched via https. The frequency depends on DR events and data is stored in an SQL database. The readings are composed of the following attributes:

Attribute	Unit
Power	W
Grid frequency	Hz
SoC	Wh
Voltage	V
Asset availability	On/Off
Power setpoint	W

Table 24 - LSP 5 data set: residential demand response assets

2. **Open data Dutch DSOs:** This data is fetched via https once per year and stored as CSV. The readings are composed of the following attributes:

Attribute	Unit
Standard year volume	KWh/y
Yearly residential production NL per postal code	KWh/y
DSO assets grid location	coordinates

Table 25 - LSP 5 data set: open data from Dutch DSOs

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

3. **Enexis MV measurements:** This data is fetched per mail on request. It is sampled every 15 minutes and stored as CSV. The readings are composed of the following attributes:

Attribute	Unit
Reactive power	kW
Active power	kW
Voltage	V
THD	%
Current	A

Table 26 - LSP 5 data set: Enexis MV measurements

3.1.6. LSP 6: Coordinating grid-owned and behind-the-meter assets management prediction for improved grid operation and near real time power market operation settlement

LSP 6 is a BD-4-DER LSP with the goal of enhancing the coordinated management of grid and third party-owned grid connected assets, and optimizing their flexibility potential with a view to support different and new business models and subsequent anticipated settlement.

The DSO's reactive (SCADA) and preventive (protection scheme) operational applications will be complemented with proactive (schedules) and corrective (mode of operation, set point modulation, etc.) operation schemes based on data analytics. Internally within the DSO (intra-stakeholder) this demands for the integration of growing data stores of disparate systems (e.g., meter data management, customer information from smart meters, outage management, etc.) with the aim of obtaining critical insights to enhance operations as well as strategic decision-making. The final expected outcome is to:

- increase efficiency and reliability of network operation via coordinated optimization and management of connected assets,
- include different possible models of consumer participation, design of rules, settlement process and assessment criteria and
- design infrastructure with reliable and near real-time metering data aggregation to support analytical tools.

The following data sets will be considered. Data that is only stored once (e.g., static data about transformer and weather stations) is omitted here.

1. **Assets Condition Data:** This data is fetched via MQTT once per minute and stored as csv. The readings are composed of the following attributes:

Attribute	Unit
Transformer Oil Temperature	°C

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

Attribute	Unit
Transformer Casing Temperature	°C

Table 27 - LSP 6 data set: assets condition data

2. **Energy Data:** This data is fetched via MQTT (DLMS/Cosem) every 15 minutes and stored as csv. The readings are composed of the following attributes:

Attribute	Unit
Transformer Station ID	/
Date_Time	YYYY-MM-DD hh:mm:ss
Active power A+	kW
Active power A-	kW
Reactive power R+	kVA
Reactive power R-	kVA
Voltage per Phase	V
Current per Phase	A
Power Factor Cos Phi	/

Table 28 - LSP 6 data set: energy data

3. **Meter Events:** This data is fetched via MQTT (DLMS/Cosem) once per event and stored as csv. The readings are composed of the following attribute:

Attribute	Unit
Maintenance	/

Table 29 - LSP 6 data set: meter events

4. **Weather Data:** This data is fetched via a web service once per hour and stored as csv. The readings are composed of the following attributes:

Attribute	Unit
Humidity	%
Ambient Temperature	°C

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

Attribute	Unit
Wind Speed	km/h
Rain	mm
Solar Radiation	W/m^2
Wind Direction	degrees
Weather Station ID	/
Date Time	YYYY-MM-DD hh:mm:ss
Weather Measurement ID	/

Table 30 - LSP 6 data set: weather data

5. **Weather Forecast Data:** This data is fetched via a web service once per hour and stored as csv. The readings are composed of the following attributes:

Attribute	Unit
Humidity Forecast	%
Ambient Temperature Forecast	°C
Wind Speed Forecast	km/h
Rain Forecast	mm
Solar Radiation Forecast	W/m^2
Location ID	/
Date of Forecast	YYYY-MM-DD
Date Time	YYYY-MM-DD hh:mm:ss
Weather Forecast ID	/

Table 31 - LSP 6 data set: weather forecast data

6. **Energy Real Time Data:** This data is fetched via MQTT (DLMS/Cosem) every minute and stored as csv. The readings are composed of the following attributes:

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

Attribute	Unit
Transformer Station ID	/
Date_Time	YYYY-MM-DD hh:mm:ss
Active power A+	kW
Active power A-	kW
Reactive power R+	kVA
Reactive power R-	kVA
Measurement ID	/
Voltage per Phase	V
Current per Phase	A
Power Factor Cos Phi	/

Table 32 - LSP 6 data set: energy real time data

7. **Weather Data Real Time:** This data is fetched via a web service once per minute and stored as csv. The readings are composed of the following attributes:

Attribute	Unit
Humidity	%
Ambient Temperature	°C
Wind Speed	km/h
Rain	mm
Solar Radiation	W/m²
Wind Direction	degrees
Weather Station ID	/
Date Time	YYYY-MM-DD hh:mm:ss
Weather Measurement ID	/

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

Table 33 - LSP 6 data set: weather data (real time)

3.1.7. LSP 7: Predictive Analytics Forecasting of storage optimized operation for improved market reserve participation

LSP7 focuses on predictive analytics forecasting of storage optimized operation for improved market reserve participation**.** The pilot aims to showcase how large-scale solar power can benefit from big data analytics and predictions applications to improve maintenance, observation, troubleshooting and monitoring operations on a large number of components that compose a PVs deployment. Operations that will be considered in the context of this pilot include:

– Maintenance of the CPV panels, including data driven scheduling of panels cleaning
– Electrical troubleshooting based on fault prediction analytics and data driven approaches
– Optimization of equipment performance and energy production output with the use of weather data

The benefits of big data technologies for LSP7 will be shown at partner ENELX, and the following datasets will be considered:

1. AC Meter readings. These data are fetched over the Modbus TCP/RTU protocol with a frequency of 1 sec and are stored in a CSV/JSON at NO SQL database. The readings are composed of the following attributes:

Attribute	Unit
Voltage Phase to Neutral	V
Current	A
Active Power	W
Apparent Power	VA
Reactive Power	Var
Demanded Power	W
Demanded Power Peak	W
Power Factor	nr
Frequency	Hz
Total Injected Active Energy	kWh
Total Withdrawn Active Energy	kWh
Partial Injected Active Energy	kWh
Partial Withdrawn Active Energy	kWh

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

Attribute	Unit
Total Injected Reactive Energy	kvarh
Total Withdrawn Reactive Energy	kvarh
Partial Injected Reactive Energy	kvarh
Partial Withdrawn Reactive Energy	kvarh
Total Apparent Energy	kVA
Partial Apparent Energy	kVA
Total Current Harmonic Distortion	A
Total Voltage Harmonic Distortion	V
Timestamp	

Table 34 - LSP 7 data set: AC meter readings

2. DC Meter readings. These data are fetched over the Modbus TCP/RTU protocol with a frequency of 1 sec and are stored in a CSV/JSON at NO SQL database. The readings are composed of the following attributes:

Attribute	Unit
Voltage	V
Minimum Voltage	V
Maximum Voltage	V
Current	A
Minimum Current	A
Maximum Current	A
Active Power	W
Minimum Active Power	W
Maximum Active Power	W
Total Active Energy	kWh
Timestamp	

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

Table 35 - LSP 7 data set: DC meter readings

3. ESS (Energy Storage System) Control data. These data are fetched over the Modbus TCP/RTU protocol with a frequency of 1 sec and are stored in a CSV/JSON at NO SQL database. The readings are composed of the following attributes:

Attribute	Unit
Inverter output power	W
Inverter output voltage	V
Inverter output current	A
Battery state of charge	%
Minimum battery state of charge value	%
ESS system current status	Binary
Total number of charge/discharge cycles	Nr
Cell voltage	V
Cell current	A
Cell temperature	°C
Module voltage	V
Module current	A
Module temperature	°C
Rack total output power	kW
Rack state of charge	%
Power Factor	nr
Timestamp	/

Table 36 - LSP 7 data set: energy storage system control data

Moreover, the ESS control system provides outputs of Logfiles with faults and warnings.

4. Load Simulator. These data are fetched over the Modbus TCP/RTU protocol with a frequency of 1 sec and are stored in a CSV/JSON at NO SQL database. The readings are composed of the following attributes:

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

Attribute	Unit
Load Phase R	%
Load Phase S	%
Load Phase T	%
Total Load (R+S+T) / #phases	%
Load Current Phase R	A
Load Current Phase S	A
Load Current Phase T	A
Load Frequency	Hz
Load Power	kW
Load Power Phase R	kVA
Load Power Phase S	kVA
Load Power Phase T	kVA
Timestamp	

Table 37 - LSP 7 data set: load simulator data

5. PV Simulator. These data are fetched over the Modbus TCP/RTU protocol with a frequency of 1 sec and are stored in a CSV/JSON at NO SQL database. The readings are composed of the following attributes:

Attribute	Unit
Average Current	A
High level of pulse waveform	A
Low level of pulse waveform	A
Maximum current value	A
Minimum current value	A

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

Attribute	Unit
Average Voltage	V
High level of pulse waveform	V
Low level of pulse waveform	V
Maximum voltage value	V
Minimum voltage value	V
Average Power	W
Maximum power value	W
Minimum power value	W
Accumulated Amp-Hours	Ah
Accumulated Watt-Hours	Wh
Timestamp	

Table 38 - LSP 7 data set: PV simulator

3.1.8. LSP 8: Predictive Demand and Generation Forecasting for Optimized Local Energy Community Management

This pilot focuses on optimized energy management on non-interconnected island electricity systems which are largely dependent on the use of imported oil for electricity generation purposes. With the advent of renewable energy sources (RES) the aim is to optimize the excellent RES potential of such islands and support actions to maximize RES integration.

There are several challenges that need to be considered, including significant load variations over the course of the year, e.g., during the winter non-touristic period a lower energy demand is observed. The energy system is such cases needs to be flexible by means of including the components of storage, DR, vector coupling and also advanced energy management relying on predictive analytics relying on big data. Aspects of energy management that will be considered focus on optimal scheduling and usage of deferrable loads at various levels:

- Optimal scheduling and usage of deferrable loads of different elasticity to achieve higher RES shares at local (island) community level
- Optimal management of island municipalities' deferrable loads
- Optimal scheduling of deferrable loads for offering distribution grid services at the island level (e.g., peak shaving)

The benefits of big data technologies and analytics for LSP8 will be shown at the Greek island of Tilos, supported by partner UNIWA. The following datasets will be considered:

1. Water Pumping Stations data.

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

Attribute	Unit
Total Active Power	W
Total Apparent Power	VA
Total Reactive Power	VAR
Voltage per Line-N	V
Power Factor per Line-N	%
Total Power Factor	%
Frequency per pumping station	Hz
Current per Line per pumping station	A

Table 39 - LSP 8 data set: water pumping stations data

2. Residential Energy Demand Data.

Attribute	Unit
Total Active Power	W
Voltage	V
Power Factor	%
Frequency	Hz

Table 40 - LSP 8 data set: residential energy demand data

3. Weather Data.

Attribute	Unit
Wind Speed	m/sec
Wind Direction	Deg
Solar Radiation	W/m^2
Temperature	^oC
Absolute Humidity	g/kg
Relative Humidity	%

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

Table 41 - LSP 8 data set: weather data

4. Kos-Kalymnos Power Generation System Data

Attribute	Unit
Theoretical Wind Power Generation per Wind Park of the system	MW
Wind Power Injected to Grid	MW
Solar Power Generation	MW
Thermal Units Commitment	MW
Total Load Demand	MW

Table 42 - LSP 8 data set: Kalymnos power generation system data

3.1.9. LSP 9: Collaborative aggregated energy generation prediction

This pilot focuses on energy communities and aims to improve the maintenance and prediction of power productions of renewable energy sources (RES). The related testbed is located in Spain and is managed by the energy community Enercoutim which has deployed a large-scale solar plant and a free-standing self-consumption system with EV charging and storage components. More specifically this LSP aims to showcase improvements of the following operations:

- Scheduling of maintenance operations to reduce unnecessary or due maintenance. Currently maintenance of assets is performed on a regular basis even if not required. With the use of data driven solutions, predictive maintenance can be optimized through proper scheduling of maintenance actions when needed.
- Predictive fault identification which will allow RES operators to perform troubleshooting before faults occur. Currently there are no means in place to anticipate faults of assets and commonly electrical troubleshooting is performed after faults occur.
- Production forecast to add value to the services provided as currently there is also no way to forecast future production of the system in an accurate way.

In terms of data availability, the following datasets will be considered:

1. Solar Lab Weather data (weather station located at Martim Longo, Portugal).

Measurements of the following attributes are provided every second. All values are numeric.

Attribute	Unit
Temperature	Celsius degrees
Humidity	%
Wind speed	m/s

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

Attribute	Unit
Direct Radiation	Wh/m^2
Global Radiation	Wh/m^2

Table 43 - LSP 9 data set:solar lab weather data

Sampling Frequency: from 10 seconds to every 10 minutes

2. Solar Lab PV installation data.

There are twelve measurements of the following attributes per hour. All values are numeric.

Attribute	Unit
Total Active Power	kW
Total Reactive Power	kVar
L1, L2 and L3 AC Voltage	V
DC Voltage	V
L1, L2 and L3 AC Current	A
L1, L2 and L3 Frequency	Hz
Battery Pac	kW
Battery SoC	%

Table 44 - LSP 9 data set: solar lab PV installation data

3. Martim Longo School Power Consumption data.

The total power consumption of the school is measured every second, in Kw (numeric).

4. Solar Demonstration Platform data.

All values are numeric.

Attribute	Unit
Aggregate Power 1	MW
Aggregate Power 2	MW
Aggregate Power 3	MW
Inverter Active Power of group 1,2 or 3	kW

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

Attribute	Unit
Inverter Reactive Power of group 1,2 or 3	kW
Inverter DC Voltage of group 1,2 or 3	V
Inverter AC Voltage of each phase in group 1,2 or 3	V
Inverter AC Current of each phase in group 1,2 or 3	A
Inverter's Frequency of each phase in group 1,2 or 3	Hz

Table 45 - LSP 9 data set: solar demonstration platform consumption data

5. Maintenance data. The following information will be considered:

Attribute	Unit/Info
DeviceId	string
CPV cleaning	#number/year
Electric board repair	#number/year
CPV string repair	#number/year
Inverter replacement	#number/year
dateObserved	date-time
elementObserved	*Enumeration of elements
value	Numeric / array of numbers
deviceObserved	string (relationship)
groupId	string (relationship)

Table 46 - LSP 9 data set: maintenance data

6. Energy forecast

Energy production forecast based on the weather conditions and other factors. Power production forecast in MW (numeric) with 15 mins resolution up to 48 hours ahead

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

Attribute	Unit/Info
DeviceId	string
dateObserved	date-time
dateForecasted	date-time
forecastMethod	*Enumerated values
powerProduction	MW

Table 47 - LSP 9 data set: energy forecasts

3.1.10. LSP 10: Improved Energy Performance Certificates Reliability

This pilot aims to provide data driven methodologies for dynamic Energy Performance Certificates that will be considering actual energy consumption data, such as real consumption and contextual information.

The pilot will run in Spain, driven by partners CARTIF and VEOLIA. The following data will be considered:

1. Open Weather data from meteorological services.

These include weather forecasts and historical data from the Spanish Meteorological Agency (AEMET), Meteostat (only historical), Weatherbit (only forecasting data), COPERNICUS SoDA (Solar radiation Data) and COPERNICUS C3S (climate change service) data.

The meteorological service to be used has not been decided, since at this stage there are things to be defined, and specific parameters to be used are not clear. Additionally, other analytics included in other pilots will need to use a weather service, and the same one could be used to avoid working on integration issues more than once.

2. Cadastre data (Open Data)

These are provided by the Spanish ministry of interiors and are provided as open data, accessible through the following URL: http://www.catastro.minhap.es/INSPIRE/buildings/ES.SDGC.BU.atom.xml

It will be compliant with the GML[10] export from the Spanish cadaster, that compliant with INSPIRE directive related with buildings[11].

The building related data coming from BU files will consist of the following attributes:

Attribute	Info/Type
gml:id	String
beginLifespanVersion,	Date
conditionOfConstruction	String
dateOfConstruction:beginning	Date
dateOfConstruction:end	Date

[10]http://www.catastro.minhap.es/documentos/formatos_intercambio/Formato%20GML%20parcela%20catastral.pdf
[11]https://inspire.ec.europa.eu/Themes/126/2892

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

Attribute	Info/Type
informationSystem	String
reference	String
localId	String
namespace	String
horizontalGeometryEstimatedAccuracy	Double (horizontalGeometryEstimatedAccuracy_uom)
horizontalGeometryEstimatedAccuracy_uom	String
horizontalGeometryReference	String
referenceGeometry	String
currentUse	String (codelist)
numberOfBuildingUnits	Integer
numberOfDwellings	Integer
documentLink	String
format	String
sourceStatus	String
officialAreaReference	String
value	Integer(value_uom)
value_uom	String -usually m2-
geometry	GML Geometry element

Table 48 - LSP 10 data set: open cadastre data

For the data coming from BU part files the following attributes will be considered:

Attribute	Unit/Info
gml:id	String
beginLifespanVersion	Date
localId	String
namespace	String
horizontalGeometryEstimatedAccuracy	Double(horizontalGeometryEstimatedAccuracy_uom)
horizontalGeometryEstimatedAccuracy_uom	String -usually m-
horizontalGeometryReference	String
referenceGeometry	String

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

Attribute	Unit/Info
numberOfFloorsAboveGround	Integer
heightBelowGround	Double (heightBelowGround_uom)
heightBelowGround_uom	String – Usually m-
numberOfFloorsBelowGround	Integer
geometry	GML Geometry element

Table 49 - LSP 10 data set: building part data

3. Energy Performance Certificates (Open data)

Provided by the regional register of Castilla y León, provided through the following URLs:

– https://datosabiertos.jcyl.es/web/jcyl/set/es/energia/certificados-eficiencia/1284543386412
– https://help.opendatasoft.com/apis/ods-search-v1/#search-api-v1

The following attributes are considered:

Attribute	Unit/Type
id	string
dateRegistry	datetime
buildingUseType	enumerated values
buildingUse	enumerated values
address	string
longitude	number
latitude	number
locality	enumerated values
province	enumerated values
ratioPrimaryConsumption	KWh/m^2·year
primaryEnergyQualification	enumerated values
co2EmissionsRatio	kgCO2e/m^2·year
emissionQualification	enumerated values
heatingDemandValue	KWh/m^2·year
heatingDemandQualification	enumerated values
coolingDemandValue	KWh/m^2·year
coolingDemandQualification	enumerated values

Table 50 - LSP 10 data set: energy performance certificates

4. Energy DataHub Castilla y León (Open Data)

Electricity consumption in centres of the autonomous administration of Castilla y León available through the following URLs:

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

- https://analisis.datosabiertos.jcyl.es/api/v1/console/datasets/1.0/search/

Attribute	Name translation	Unit / Type
Organismo / consejería	Organism / counsil	string
Centro gestor	Management center	string
Id opte centro de gasto	Spending center id	integer
Centro de gasto	Spending center	string
Tipo de centro de consumo	Type of consumer center	string
Id opte centro de consumo	Consumer center id	integer
Centro de consumo	Consumer center	string
Cups electricidad	Electricity cups	string
Tipo de energía	Type of energy	string
Distribuidora eléctrica	Electric distributor	string
Tipo suministro eléctrico	Type of electrical supply	string
Tipo de punto de medida	Type of measuring point	string
Tarifa eléctrica	Electric rate	string
P1 (kw) potencia contratada	P1 (kw) contracted power	kW
P2 (kw) potencia contratada	P2 (kw) contracted power	kW
P3 (kw) potencia contratada	P3 (kw) contracted power	kW
P4 (kw) potencia contratada	P4 (kw) contracted power	kW
P5 (kw) potencia contratada	P5 (kw) contracted power	kW
P6 (kw) potencia contratada	P6 (kw) contracted power	kW
Año	Year	string
Mes	Month	string
Consumo mensual energía activa total (kwh)	Monthly consumption of total active energy (kwh)	kWh
G.d. En base 20	G.d. In base 20	integer
G.d. En base 26	G.d. In base 26	integer
Tipo de centro a nivel de administración autonómica	Type of center at the regional administration level	string
Código municipio dirección	Municipal code address	integer
Municipio dirección	Municipality address	string
Coordenada x longitud	Coordinate x length	decimal
Coordenada y latitud	Coordinate y latitude	decimal

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

Attribute	Name translation	Unit / Type
Posición	Position	geo_point_2d
Fecha	Date	date

Table 51 - LSP 10 data set: energy consumption data

Diesel consumption in Education and Health centres of the autonomous administration of Castilla y León:

Attribute	Name translation	Date Type
Organismo / consejería	Organism / counsil	string
Centro gestor	Management center	string
Id opte centro de gasto	Spending centre id	string
Centro de gasto	Spending centre	integer
Tipo de centro de consumo	Type of consumer centre	string
Id opte centro de consumo	Consumer centre id	integer
Centro de consumo	Consumer centre	string
Tipo de energía	Type of energy	string
Año	Year	string
Mes	Month	string
Consumo mensual total gsl (m3) gasóleo a	Monthly consumption total gsl (m3) a diesel	decimal
Consumo mensual total gsl (m3) gasóleo b	Monthly consumption total gsl (m3) b diesel	Decimal
Consumo mensual total gsl (m3) gasóleo c	Monthly consumption total gsl (m3) c diesel	decimal
G.d. En base 20	G.d. In base 20	Integer
G.d. En base 26	G.d. In base 26	integer
Tipo de centro a nivel de administración autonómica	Type of centre at the regional administration level	string
Código municipio dirección	Municipal code address	string
Municipio dirección	Municipality address	string
Coordenada x longitud	Coordinate x length	decimal
Coordenada y latitud	Coordinate y latitude	decimal
Posición	Position	geo_point_2d
Fecha	Date	date

Table 52 - LSP 10 data set: diesel consumption

Gas consumption in centres of the autonomous administration of Castilla y León:

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

Attribute	Name translation	Data Type
Organismo / consejería	Organism / counsil	string
Centro gestor	Management centre	string
Id opte centro de gasto	Spending centre id	Integer
Centro de gasto	Spending centre	string
Tipo de centro de consumo	Type of consumer centre	string
Id opte centro de consumo	Consumer centre id	integer
Centro de consumo	Consumer centre	string
Cups gas natural	Natural gas cups	string
Tipo de energía	Type of energy	string
Distribuidora gas natural	Gas natural distributor	string
Uso gas natural	Gas natural use	string
Tarifa gas natural	Gas natural rate	decimal
Año	Year	string
Mes	Month	string
Consumo mensual energía activa total (kwh)	Monthly consumption of total active energy (kwh)	Integer
G.d. En base 20	G.d. In base 20	Integer
G.d. En base 26	G.d. In base 26	integer
Tipo de centro a nivel de administración autonómica	Type of centre at the regional administration level	string
Código municipio dirección	Municipal code address	integer
Municipio dirección	Municipality address	string
Coordenada x longitud	Coordinate x length	Decimal
Coordenada y latitud	Coordinate y latitude	Decimal
Posición	Position	geo_point_2d
Fecha	Date	date

Table 53 - LSP 10 data set: gas consumption

Hourly consumption in hospitals:

Attribute	Name translation	Data Type
Organismo / consejería	Organism / counsil	string
Centro gestor	Management centre	string
Id opte centro de gasto	Spending centre id	Integer

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

Attribute	Name translation	Data Type
Centro de gasto	Spending centre	string
Tipo de centro de consumo	Type of consumer centre	string
Id opte centro de consumo	Consumer centre id	integer
Centro de consumo	Consumer centre	string
Cups electricidad	Electricity cups	string
Fecha de lectura	Reading date	date
Año	Year	integer
Mes	Month	integer
Día	Day	integer
Hora	Hour	string
Día de la semana	Day of the week	string
Tipo de día	Type of day	Integer
Estación	Season	string
Tarifa	Rate	string
Periodo	Period	integer
Consumo horario energía activa (kwh)	Hourly active energy consumption (kwh)	integer
Precio energía activa	Active energy price	decimal
Grado 20 diarios	Grade 20 daily	Integer
Grado 26 diarios	Grade 26 daily	integer
Tipo de centro a nivel de administración autonómica	Type of centre at the regional administration level	String
Código municipio dirección	Municipal code address	integer
Municipio dirección	Municipality address	String
Coordenada x longitud	Coordinate x length	decimal
Coordenada y latitud	Coordinate and latitude	decimal
Posición	Position	Geo_point_2d
Fecha	Date	date
Gasto_lectura	Expense_read	decimal
N_dia_semana	N_day_week	integer

Table 54 - LSP 10 data set: consumption in hospitals

List of all energy consumption centres of the autonomous administration of Castilla y León, including data such as expenditure centres, geoposition and cadastral reference.

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

Attribute	Name translation	Data Type
Organismo / consejería	Organism / counsel	string
Centro gestor	Management centre	string
Id opte centro de gasto	Spending centre id	integer
Centro de gasto	Spending centre	string
Código interno	Internal code	string
Id opte centro de consumo	Consumer centre id	integer
Centro de consumo	Consumer centre	string
Tipo de centro de consumo	Type of consumer centre	string
C.I.F.	C.I.F.	string
Dirección	Address	integer
Superficie construida	Built surface	integer
Ocupación	Occupation	integer
Cups E	Code of supply point (Electricity)	integer
Cups GN	Code of supply point (Natural Gas)	integer
GSL	Code of supply point (Diesel)	integer
Tipo de centro a nivel de administración autonómica	Type of centre at the regional administration level	string
Código municipio dirección	Municipal code address	string
Municipio dirección	Municipality address	string
Coordenada x longitud	Coordinate x length	Decimal
Coordenada y latitud	Coordinate y latitude	decimal
Referencia catastral 1	Cadastral reference 1	string
Referencia catastral 2	Cadastral reference 2	string
Referencia catastral 3	Cadastral reference 3	string
Referencia catastral 4	Cadastral reference 4	string
Referencia catastral 5	Cadastral reference 5	string
Referencia catastral 6	Cadastral reference 6	string
Referencia catastral 7	Cadastral reference 7	string
Posición	Position	Geo_point_2d

Table 55 - LSP 10 data set: energy consumption centres data

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

5. Building audits*

Final Energy saving; CO_2 emissions avoided, estimated investments, R.O.I., ECM energy savings, certificate management for buildings, provided through the following URL:

https://analisis.datosabiertos.jcyl.es/explore/dataset/auditorias-energeticas-en-castilla-y-leon/table/?sort=-fecha_auditoria

Attribute	Data Type
Province	string
InvestmentsForImprovements	String: Eumerated value: León, Zamora, Salamanca, Valladolid, Palencia, Soria, Segovia, Avila, Burgos
co2EmissionsAvoided	datetime
ReturnOfInvestment	string
dataObserved	Date-time
energySavingsImprovements	Number
hasEnergyManagementCertification	Boolean
finalEnergyConsumption	Number

Table 56 - LSP 10 data set: building audit data

9. Occupant behaviour

These data will include Temperature, humidity, CO2 per dwelling, ºC, %, ppm.

10. HEMS / BEMS / DEMS

Energy consumption for the heating and/or DHW kWh

Main variables of the facilities: temperatures of the circuit, pressures, fuel consumption

Example from Sierra Elvira:

Attribute	Name translation	Unit / Data Type	
Tª exterior	Outdoor temperature	decimal	ºC
Temp amb aula music pta segun	Classroom temperature music room second door	decimal	ºC
Temp ambte aula 3c pta tercera	Classroom temperature room 3c third	decimal	ºC
Temp ambte aula 5c pta tercera	Classroom temperature room 5c third	decimal	ºC
Radiacion solar	Solar radiation	decimal	W/m²

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

Attribute	Name translation	Unit / Data Type	
Peso del silo	Silo weight	decimal	
Temp impulsion caldera	Boiler delivery temp	decimal	ºC
Sonda presion	Pressure probe	decimal	
Temp. Ambte salon comed pta 1	Living room temperature door 1	decimal	ºC
Temp. Ambte comed 3 años pta 1	Living room temperature door 1 3 years	decimal	ºC
Temp. Ambte aula inform pta 2	Inform classroom temperature door 2	decimal	ºC
Temp. Ambte aula 3ºb planta 2	Classroom temperature 3ºb floor 2	decimal	ºC
Temp. Ambte aula 5ºb planta 3	Classroom temperature 5ºb floor 2	decimal	ºC
Temp. Ambte aula 6ºb planta 3	Classroom temperature 6ºb floor 3	decimal	ºC
Tª impulsion biblioteca	Library delivery temperature	decimal	ºC
Tª retorno biblioteca	Library return temperature	decimal	ºC
Temp retorno caldera	Boiler return temp	decimal	ºC
Tª retorno zona 1	Return temperature zone 1	decimal	ºC
Tª retorno zona 2	Return temperature zone 2	decimal	ºC
Presion diferenc circuito 1	Pressure difference circuit 1	Decimal	
Presion diferenc circuito 2	Pressure difference circuit 2	Decimal	
Contador energia cto 1	Energy counter cto 1	Decimal	kWh
Contador energia cto 2	Energy meter cto 2	decimal	kWh
Tª impulsion zona 1	Delivery temperature zone 1	decimal	ºC
Tª impulsion zona 2	Delivery temperature zone 2	decimal	ºC

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

Attribute	Name translation	Unit / Data Type	
Temp amb comed 4 años pta prim	Living room temperature 4 years old first door	decimal	ºC
Temp amb secretaria pta primer	Secretary temperature first door	decimal	ºC
Temp ambte aula 4a pta segunda	Room 4a temperature second door	decimal	ºC
Contador energia (1 hora)	Energy counter (1 hour)	decimal	kWh

Table 57 - LSP 10 data set: data from facilities (example)

Main variables of the facilities: temperatures of the circuit, pressures, fuel consumption

11. Invoices

Depending on the final buildings to be included, the availability of this information will be checked.

Periodical energy consumption of the facility kWh??

12. Comparison of results vs EPC

Correct outstanding EPC values. It is expected improvement in EPC management (CO_2, demand for heating, cooling, consumption of primary energy without renewables)

13. Future scenarios of temperature evolution

Climate change impact in the face of changes in the meteorological pattern and therefore measures to improve the efficiency of buildings.

3.1.11. LSP 11: Weather-enhanced building thermal comfort prediction to enable energy efficiency reliable forecasting

LSP 11 aims at showing how big data can be used for improving energy efficiency in buildings and supporting near real time comfort-enhanced decision-making by building or district operators.

The aim is to consider occupants' thermal comfort, combine it with building/district energy consumption data and weather information in order to generate thermal comfort predictions at building level while optimizing energy efficiency and ensuring thermal comfort. For example, benefits can be gained by linking HVAC set points and actions on controllable loads with thermal comfort predictions.

The pilot will be deployed in the San Cugat region in Spain. The following data will be considered:

1. Open Weather data from meteorological services.

These include weather forecasts and historical data from the Spanish Meteorological Agency (AEMET), Meteostat (only historical), Weatherbit (only forecasting data), COPERNICUS SoDA and COPERNICUS C3S data.

2. Cadastre data (Open Data)

These are provided by the Spanish ministry of interiors and are provided as open data, accessible through the following URL: http://www.catastro.minhap.es/INSPIRE/buildings/ES.SDGC.BU.atom.xml

3. LiDAR data

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

Classification (unassigned, ground, low vegetation, medium vegetation, high vegetation, building, noise) as provided by http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=LIDAR

4. BEMS

Energy consumption, temperatures, pressures, flow rates, etc.

5. Invoices

3.1.12. LSP 12: Predictive Analytics for Energy Efficiency Investments de-risking

In the current situation a large number of decision makers have only limited information on the actual energy consumption of buildings. Indeed, actual data shows that EPC is not very reliable in assessment of energy performance, especially when it comes to deciding on the effectiveness of financial investments in these buildings.

Many public institutions and project developers, in particular when there is the support of an ESCO company, lack information and knowledge about the implementation of many energy efficient projects, their economic justification and historical energy data consumption.

The pilot aims to make use of big data for predicting the value of energy efficiency investments. Currently it is common that the actual consumption in buildings and related energy performance calculations rely of energy performance certificates which do not consist a reliable source of information. In order to implement new energy efficiency projects detailed knowledge on the historical and real energy consumption is needed, which can be used to provide justification on the corresponding investments and estimate the related efficiencies before moving on the implementation aspects.

The pilot will show provide investors / financiers and project developers the means to:

- Evaluate key performance indicators (KPIs) for energy efficiency projects;
- Explore past and future projects (project history, synopsis, projects/companies more attractive to investors);
- Make a profound risk measurement and evaluation (implement key indicators, such as debt service coverage ratio, integrate a cross-check for legal requirements of environmental laws and regulations).

The pilot will be led by partner LEIF and the following data will be considered:

- Heat energy consumption through Metbox[12] sensors
- Temperature, carbon dioxide, humidity with Aranet[13] sensors
- Electricity consumption with Sadales tīkls[14] sensors.

3.2. Overview of data clusters

The available and envisaged LSP data can be grouped in four (4) main data clusters as shown in Figure 5, which also depicts their relevance of the BD4NRG LSP clusters (BD-4-NET; BD-4-DER; BD-4-ENEF).

Energy production data are provided as aggregate data for the energy system or for 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, are stored in SQL

[12] http://metbox.com/
[13] https://aranet.com/
[14] https://www.sadalestikls.lv/en/homepage/

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

databases or files in the form of csv/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 for the different assets of the grid, including power transformers and transmission towers, the location of these assets and related incidents / work order for their maintenance. Similarly, to energy production data, the data related to the transmission / distribution system can be extracted by SCADA systems and are stored in SQL databases or files in the form of csv/json. Transmission and distribution system data are relevant for the BD-4-NET (LSP1, LSP2) and BD-4-DER (LSP4, LSP5, LSP6) clusters.

Energy consumption data are quite diverse and concern building / residential consumption, data related to electric vehicles charging, large industrial loads (water pumping stations) and load simulation data. Such data are provided by different energy management systems (such as HEMs, BEMs, DEMs) or are stored in SQL databases/ csv or json files. Energy consumption data are relevant for all pilot clusters: BD-4-NET (LSP3); BD-4-DER (LSP6, LSP8, LSP9); BD-4-ENEF (LSP10, LSP11, LSP12).

Environmental and contextual data are also provided and will be used in the BD4NRG analytics services. Such data are related i) to building assets (e.g., loads/appliances that can be used in demand-response scenarios), ii) to measurements for the indoor environment in buildings, iii) to environmental information, mainly focusing on local and area wide weather data, iv) Administrative data from public services. Such data are mainly fetched through open APIs that are provided by the data owners (e.g., weather services or public administration services). Such data are relevant for all pilot clusters. In more details 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 environment is 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).

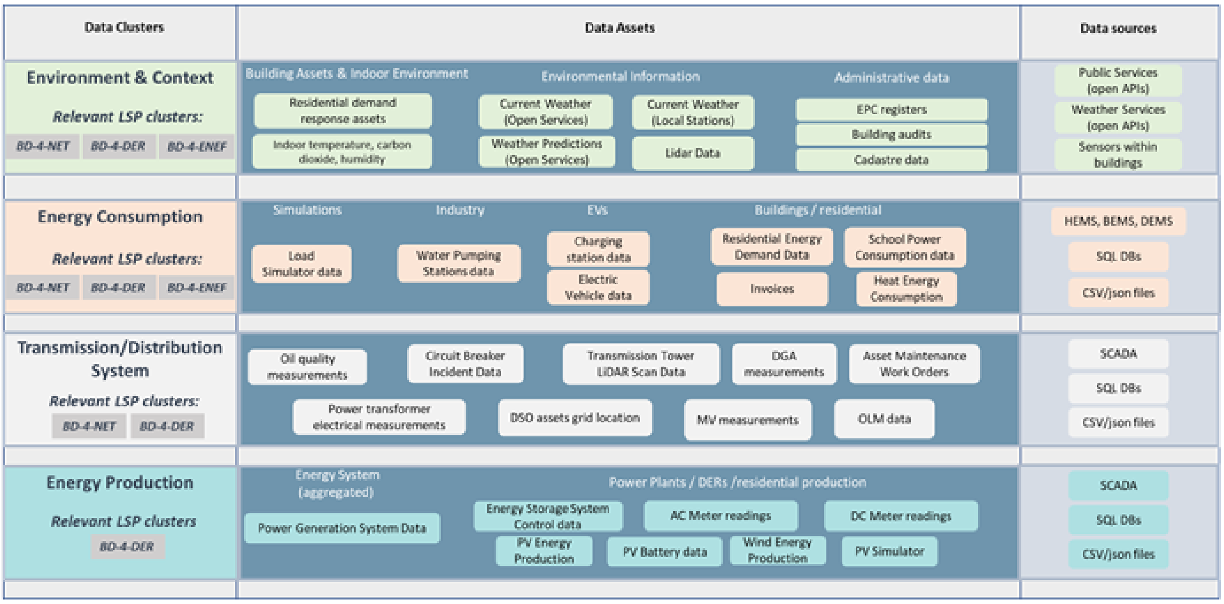

Figure 5: Overview of Data Clusters relevant for BD4NRG. The project LSPs have identified a number of data types from corresponding data sources that will be used by the analytics engines in order to create value.

3.3. Data shared between LSPs

This subchapter presents how the defined data clusters apply to BD4NRG LSPs and identifies common data sets.

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

An index of the data assets availability per LSP is provided in the following table which summarizes the information per LSP and shows the data assets that will be used per pilot (note that Annex 5.4 provides a more detailed view of the same information, listing the data assets per LSP).

Data cluster	Data asset	Large Scale Pilot											
		1	2	3	4	5	6	7	8	9	10	11	12
Energy Production	AC Meter Readings						X	X					
	DC Meter Readings							X					
	ESS Control data							X					
	Load Simulator							X					
	PV Simulator							X					
	Kos-Kalymnos Power Generation System Data								X				
	Solar Lab PV Installation Data									X			
	Solar Demonstration Platform data									X			
	Open Data Dutch DSOs					X							
Energy Grid Assets/Devices	SCADA Data		X		X								
	Transmission Tower & OHL LiDAR Scan data	X											
	Circuit Breaker Incident Data	X											
	Technical Data		X										
	OLM Data		X										
	DGA Measurements		X										
	Oil quality Measurements		X										
	Power Transformer electrical Measurements		X										

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

Data cluster	**Data asset**	**Large Scale Pilot**											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
	Work orders		X										
	Meter Events			X			X						
	Assets Condition Data						X						
	Enexis MV measurements					X							
Energy Consumption	Charging Station data			X									
	EV data			X									
	Water Pumping Station data								X				
	Residential Energy Demand Data								X				
	School Power Consumption Data									X			
	BIM										X		
	HEMS										X		
	BEMS										X	X	
	DEMS										X		
	Invoices										X		
	Heat energy consumption invoices												X
	Electricity Consumption						X						X
Building assets and Indoor Environment	Residential demand response assets					X							
	Occupant behaviour										X		
	Current Weather Data			X			X	X					

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

Data cluster	Data asset	Large Scale Pilot											
		1	2	3	4	5	6	7	8	9	10	11	12
Weather and Environmental Information	Solar Lab Weather Data (Local Weather Station)									X			
	Open Weather Data from meteorological services						X				X	X	
	Lidar Data											X	
	Temperature, CO_2, Humidity												X
	Weather forecast data				X		X						
	Weather reanalysis data				X								
Administrative Data / Public Services	Cadastre data										X	X	
	EPC Registers										X		
	Energy DataHub Castilla y León										X		
	Building audits										X		
	License files											X	
	Building EPC											X	

Table 58 - occurrences of data sets in LSPs

3.4. Conclusion

Evaluation of the data provided by BD4NRG LSPs showed that very different data assets are collected across the pilots' landscape. In order to be able to manage the inherent variety and variability of the available data assets, as well as to provide a comprehensive overview for a better understanding, the types of data assets were organized into clusters that provide structure to this diverse situation and makes it possible to gain an overview of the LSP's alignment to each cluster.

The identified data assets fall into four main categories, namely Energy production data, Data coming from the transmission / distribution system, Energy consumption data and Environmental and contextual data. However, the specific data assets within each category are quite diverse in terms of source and data type, as shown in Section 2.3 and very few individual data assets are actually common across more than one LSP.

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

Additionally, one needs to keep in mind that all pilots are still in their very early stages. This means that no complete information is available at the time of this deliverable's submission, and the existing information may be subject to changes later in the project's lifetime.

All these aspects result in the conclusion that no single data model will suffice for ensuring interoperability, and the combination of several will be necessary to cover the diverse data sets being handled within BD4NRG.

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

4 Useful Data Models

This chapter compiles the most relevant standardizations available and relevant for the project. In the introduction the concept of agile standardization is presented. This concept that will be explored in the LSPs of the project, since it allows quick standardization of data models when there is not an available standard, or when the available standards do not provide solutions for the requirements of the LSP.

4.1. Agile Standardization Approach

Comparison between classical standardization process and Agile standardization

Concept	Classic	Agile
Standardization body	Stable and fully dedicated entity → DIN	Unstable or not fully dedicated → FIWARE foundation
Standardization groups	Extended and Balanced members from relevant and interested actors	Interested users and developers with expertise on the field
Consensus mechanisms	Global Reviews by participants	By contribution and Benevolent dictator
Prestige of standards	By source entity	By use of the results
Main advantages	Low-biased standards. Coherence with former standardizations Predictable reviews Funding based on standard costs and members fees	Standard created together with implementation Evolution by use Quick reviews Low cost
Disadvantages	Creation of theoretical standards (never implemented) Slow reviews	Standardization hardly supports costs Potential biased standards Low barriers to Potential competitors

Table 59- classic vs. agile standardization

Outcomes from agile standardization processes can be an input for classical standardization, so both approaches should not be seen as alternatives but as complementary. But it is also true that not everything that can be an outcome of an agile standardization needs to become a formal standard.

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

4.2. Smart Data Models

This project is using, when applicable, an agile standardization process based on the Smart Data Models[15] initiative, which follows most of the principles of an agile standardization process. This agile initiative has these principles:

- Open: Using open licensing to allow extensive use and adoption (allows free use, free modification, and free sharing for all forever.
- Community: Data models are provided by the users not created by the initiative
- Coherence: The initiative checks that the defined data model defines properties that are consistent (same name means same data type in it)
- Evolution: Allow quick evolution of data model without heavy / slow review processes
- NGSI compatible: Compatible with NGSI standard but also beyond that standard.
- Experienced: Based on real case scenarios (so data models are tested in real case scenarios before accepted)
- Adopting what is adopted: Based on widely adopted standards (including ontologies and international schemas (i.e., schema.org)
- Git based: Based on git platform and GitHub as development frontend

The Smart Data Models initiative is an agile standardization initiative that compiles more than 400 data models for Energy infrastructure. Additionally, it also compiles data models for those domains related to the energy management, production, and distribution. Specifically, weather, environment, etc.

The 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[16],

It is a collaborative effort, so it's possible to evolve the model and to suggest modifications and new data models. Data models are grouped into subjects.

Some of the most relevant subjects for the BD4NRG project are listed below:

- **Energy**. https://github.com/smart-data-models/dataModel.Energy
- **EnergyCIM**. https://github.com/smart-data-models/dataModel.EnergyCIM
- **GreenEnergy**. https://github.com/smart-data-models/dataModel.GreenEnergy
- **Battery**. https://github.com/smart-data-models/dataModel.Battery
- **Weather**. https://github.com/smart-data-models/dataModel.Weather
- **Environment**. https://github.com/smart-data-models/dataModel.Environment

This initiative is aligned with several of the underlying principles of the European Infrastructure Framework:

- **Underlying principle 2: Openness**. All data models are open licensed and based on open standards (Recommendation 1). And they are designed to work with open-source platforms like FIWARE platform which include CEF context broker one of the building blocks of the European Union services (Recommendation 2). All their specifications are also open (Recommendation 4)
- **Underlying principle 3: Transparency**. The process to contribute is openly available (Recommendation 4, Enabling visibility). And he works in progress are also visible to allow public collaboration.
- **Underlying principle 4: Reusability.** The initiative is not defining standards but providing agile standardization to those based on actual standards (Recommendation 6) and, only in the case that there

[15] https://smartdatamodels.org
[16] https://github.com/smart-data-models

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

is not any open standardization is providing a quick solution providing three elements. 1) a technical schema capable of validations of compliant payloads, 2) a 4-Languages translated specifications and 3) some actual examples and 4) Additionally some exports for use into simple platforms (csv) or digital twins.

- **Underlying principle 5: technological neutrality and data portability**. The data models gathered in Smart Data Models count with the ability to be used with all their defined attributes, with part of them, or for particular purposes be extended with their own attributes for every user. It prevents any kind of vendor lock-in (Recommendation 8). Besides this the initiative counts with an active policy of exporting the data models to other platforms and formats (csv, geo-json feature and DTDL already available and SQL is on the queue of coming features) (Recommendation 9)
- **Underlying principle 6: User-centricity**. The data models are in constant evolution and any user can place an issue (request for information or clarification) of a pull request (suggestion for a change). In fact, the link to make such requests are in the README at the front of any of the data models (Recommendation 11). There is a weekly open call for clarification with any user willing to participate. See the calendar (Recommendation 12).
- **Underlying principle 9: multilingualism**. The data models' specifications are available in 4 languages (EN, DE, FR, SP) and some other are in the queue (IT, JP) (Recommendation 16).
- **Underlying principle 11: preservation of information.** All the data models are kept in git platform, which keeps all former versions of each data model, its specifications, and its examples (Recommendation 18).

4.3. SAREF family

Originally, the Smart Applications REFerence ontology (SAREF) is published by ETSI and intended to enable interoperability between solutions from different providers and among various activity sectors on the Internet of Things (IoT), thus contributing to the development of the global digital market.

SAREF specifies recurring core concepts in smart applications, the main relationships between these concepts, and axioms to constrain the usage of these concepts and relationships. The creators of SAREF have followed a few design principles that are in line with those specified by the European Interoperability Framework (as listed in the previous section), namely: principles of **reuse and alignment** of concepts and relationships that are defined in existing assets, **modularity** to allow separation and recombination of different parts of the ontology depending on specific needs, **extensibility** to allow further growth of the ontology, and **maintainability** to facilitate the process of identifying and correcting defects, accommodate new requirements, and cope with changes in (parts of) SAREF.[17]

Currently (July 2021), it comprehends these ontologies:

1. SAREF: the core Smart Applications REFerence ontology
2. SAREF4SYST: a generic extension of SAREF that defines patterns for Systems, Connections, and Connection Points

Extensions for domains:

1. SAREF4ENER: extension for the Energy domain
2. SAREF4ENVI: extension for the Environment domain
3. SAREF4BLDG: extension for the Building domain
4. SAREF4CITY: extension for the Smart Cities domain
5. SAREF4INMA: extension for the Industry and Manufacturing domains

[17]Source: ETSI TS 103 264

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

6. SAREF4AGRI: extension for the Smart Agriculture and Food Chain domains
7. SAREF4AUTO: extension for the Automotive domain (under development)
8. SAREF4EHAW: extension for the eHealth/Ageing-well domain
9. SAREF4WEAR: extension for the Wearables domain
10. SAREF4WATR: extension for the Water domain

The continuous addition of more extensions turns SAREF into the standards family that enables better integration of semantic data from and across various vertical domains in the IoT world.

For the sake of BD4NRG the most relevant is #1 SAREF4ENER. The list of the elements defined in this ontology can be found in the Annex. Other domain-specific extensions could be of use, depending on the domain of interest (see the Data Clusters of chapter 3). For example, LSPs from the BD-4-ENEF category might benefit from the SAREF4BLDG (Building) extension.

4.4. Semantic Sensor Network ontology

Semantic Sensor Network Ontology, SSN, is a W3C Recommendation published on 19 October 2017. It is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. SSN follows a horizontal and vertical modularization architecture by including a lightweight but self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties. With their different scope and different degrees of axiomatization, SSN and SOSA are able to support a wide range of applications and use cases, including satellite imagery, large-scale scientific monitoring, industrial and household infrastructures, social sensing, citizen science, observation-driven ontology engineering, and the Web of Things.

These images summarize most of the elements defined in this ontology.

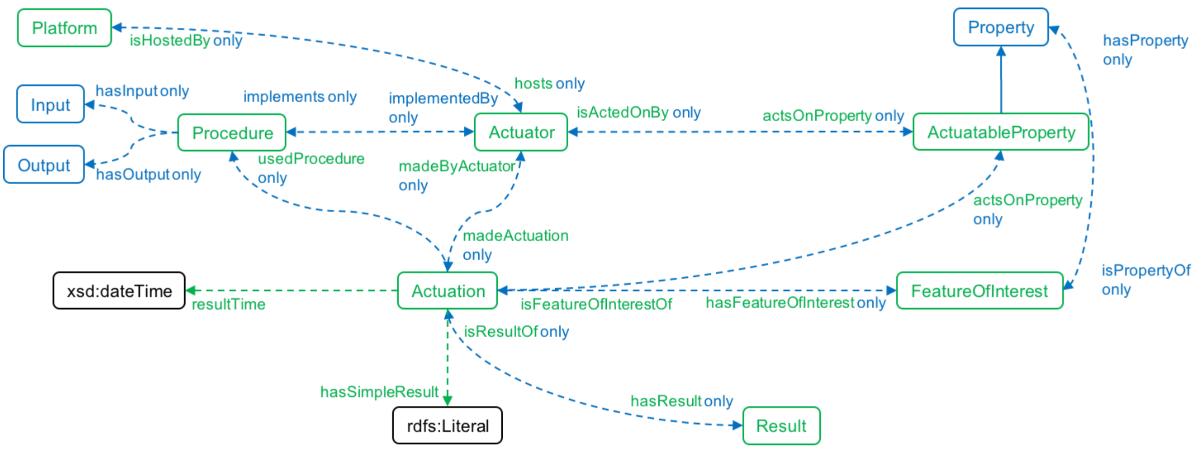

Figure 6: SSN definition of actuation.

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

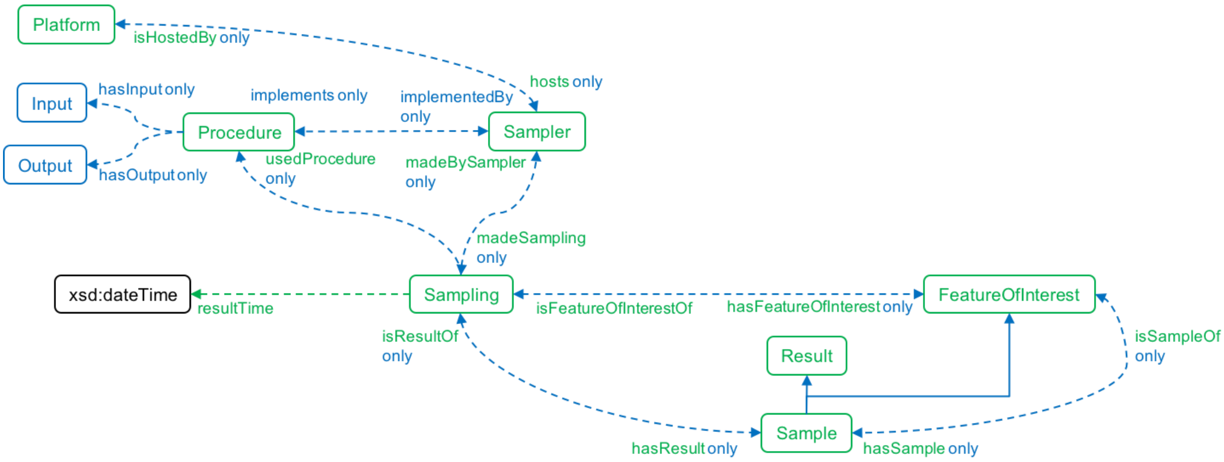

Figure 7: SSN definition of sampling.

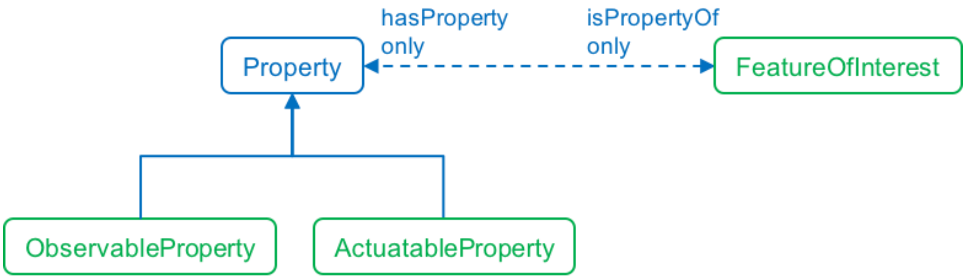

Figure 8: SSN definition of Features of Interest and Properties.

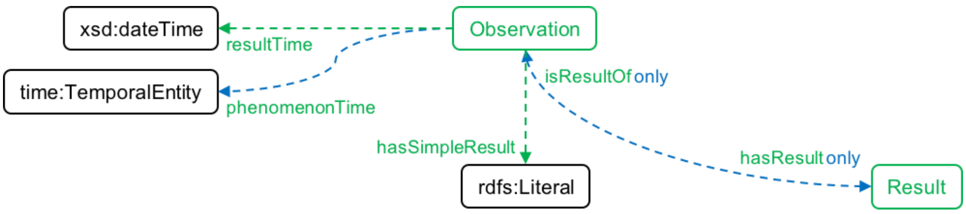

Figure 9: SSN definition of results.

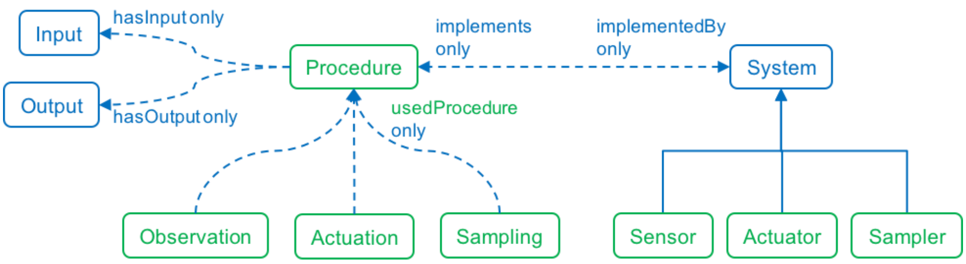

Figure 10: SSN definition of procedures

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

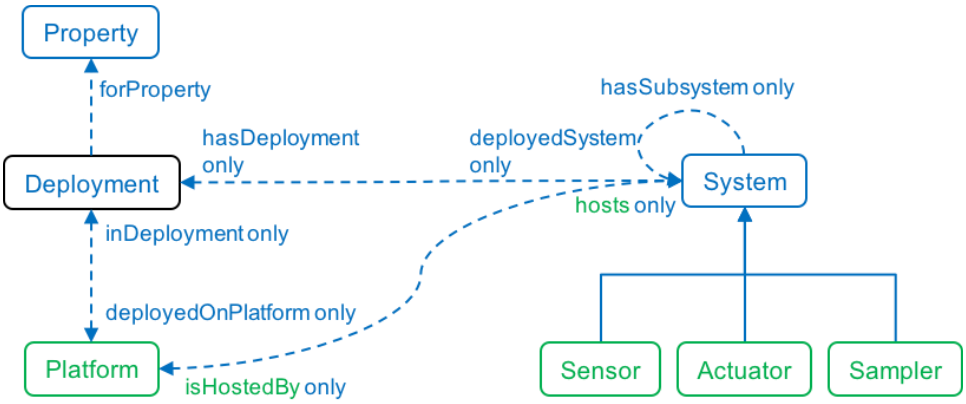

Figure 11: SSN definition of Systems and their Deployment.

4.5. EFI (S2)

The Energy Flexibility Interface (EFI) is specifically designed as a standard communication protocol between smart devices and Demand-Side Management (DSM) solutions. It is a standard that can be used to describe and control the energy flexibility of smart devices. EFI was developed to deal with interoperability issues encountered while experimenting and researching energy flexibility in field trials. It is a key enabling technology for the widespread deployment and adoption of Demand-Site Management to exploit energy flexibility. The interface specification of EFI is open and freely available.

In the meantime, EFI has had a successor called 'S2'. S2 is developed as European standard within CEN/CENELEC. It is described in the EN50491-12 series, of which the general architecture part is already finished (EN50491-12-1). The second part, EN50491-12-2, is expected to be published at the end of 2021. S2 is continuously being further improved through collaboration with industry.

The S2 interface supports a clear separation of concerns between devices and CEMs by focusing on a generic description of energy flexibility. Instead of standardizing the communication of the utilization system, or the device parameters, EFI/S2 standardizes the flexibility behaviour. It doesn't make any assumptions about the working of the utilization system or the device itself. The flexibility capabilities that smart devices provide can be mapped onto this generic description of flexibility, while it also does not impose any limitations on CEM algorithms to exploit this flexibility in a specific way.

This focus on flexibility behaviour is one of S2's greatest strength, since it is not possible to create an exhaustive list of all possible smart devices and all possible energy management strategies for the expected lifetime of the S2 specification, let alone all possible combinations of devices and CEMs.

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

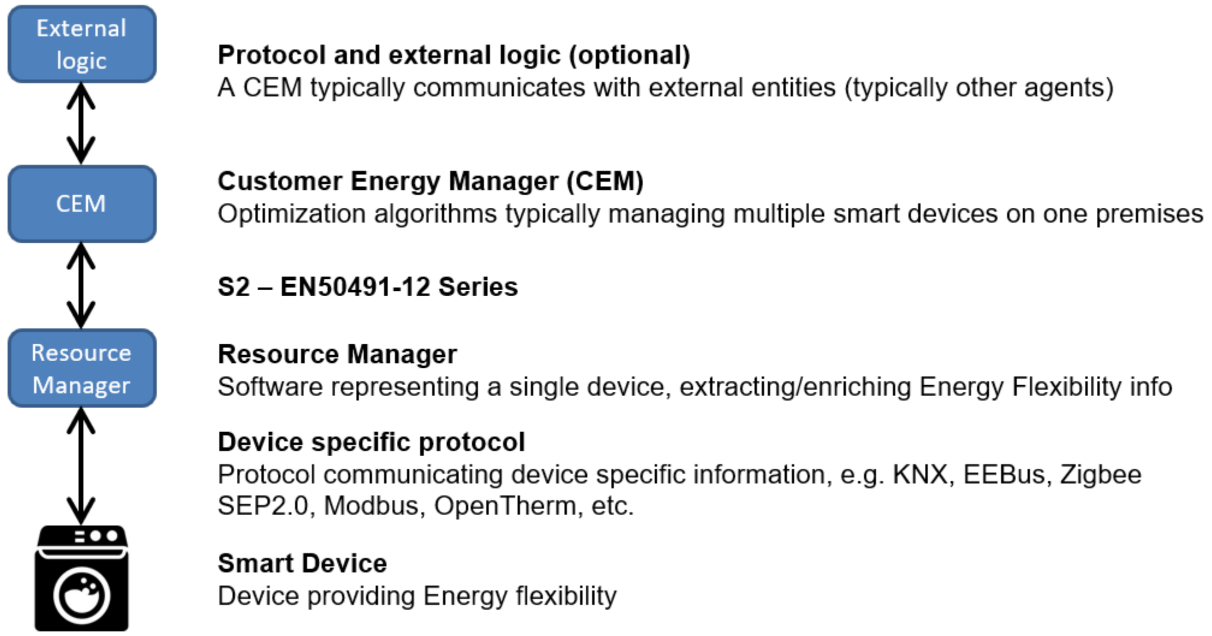

Figure 12: S2 architecture

The design of S2 is based on the observation that there's only a limited number of types of energy flexibility. Well known examples are load curtailment, load shifting, and using assets that have some kind of energy storage. There are a few more but not many. However, these flexibility patterns typically do not occur in isolation, but influence each other. These interdependencies and constraints are modelled in S2. In addition, each device can provide power measurements and power forecasts.

4.6. SARGON and various others

In the preceding subsections a number of ontologies were presented. We now approach a paradox. According to the BRIDGE Data Management working group[18], the proliferation of standards is a major obstacle to the advancement of semantic interoperability in the energy domain, and in general. Making more standards is said to only exacerbate the problem. Nonetheless, in this subsection a selection of notable ontologies and their characteristics are described.

Firstly, SARGON[19] (Smart Energy Domain Ontology) drives interoperability due to it being based on several industry standards, real uses cases and standardization efforts in the IoT world. Extending the aforementioned SAREF it keeps the modularity that allows future extensions to close any gaps such as missing information models, insufficient descriptions, and levels of object properties and relationships.[20] The initial SAREF4ENER extension to SAREF only covers a limited set of use cases such being the automation of medium voltage distribution grids (including hybrid AC/DC networks), the inclusion of Phasor Measurement Units (PMUs), the control of energy provision in buildings (gas, electricity, hot drinking water...), and the consideration of

[18] Data Management Working Group, July 2019. *Main findings and recommendations.* Downloaded at https://www.h2020-bridge.eu/wp-content/uploads/2018/06/BRIDGE-Data-Management-WG-Findings-and-Reco-July-2019.pdf
[19] Haghgoo, M., Sychev, I., Monti, A., & Fitzek, F. H. P. (2020). SARGON – Smart energy domain ontology. *IET Smart Cities*, 2(4), 191–198. https://doi.org/10.1049/iet-smc.2020.0049
[20] Blogpost on https://ideasforum.org/the-sargon-smart-energy-domain-ontology-ontology

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

Distributed Energy Resources (DER) and prosumers in city districts regarded as building agglomerations. Please refer to the Table 60 for an overview of the additions of SARGON compared to SAREF.[21]

Supports	SARGON	Original SAREF
HVAC Systems	**Yes**	No
Smart Grid	**Yes**	No
Lighting Systems	**Yes**	No
Electrical Systems	**Yes**	No
Spatial Information	**Yes**	No
Sensors Systems	**Yes**	**Yes**
Control Relationships	**Yes**	No
Sequential Relationships	**Yes**	No
Formal Definitions	**Yes**	**Yes**

Table 60 - A comparison: SARGON extension to SAREF versus vanilla SAREF features.

This additional support stems from identifying the required information of various use cases in the context of an examination of preceding models including SAREF4ENER and unifying them into a single ontology network as previously done e.g., in the OEMA ontology network. Simply put this is done by taking these ontologies that jointly cover the energy domain as much as possible and eliminating areas where they overlap by means of (among others) renaming ontology elements, creating new classes to relate to reused concepts, and relocating knowledge from one ontology to another. These and additional rule-based steps can be performed manually or semi-automatically as in SARGON. An assortment of tools that detect design-time errors, as well as validate the fulfilment of requirements and consistency were applied in the development SARGON.

Other than SAREF4ENER the preceding body of work that was merged into SARGON and OEMA includes ThinkHome which introduces artificial intelligence to control a smart home digital ecosystem which includes its own ontology. BOnSAI[22] introduces an ontology for service-oriented smart buildings. A fourth ontology that was examined is ProSGV3[23] which is a prosumer-oriented ontology that supports among others water heating devices, cleaning devices, appliances, charging devices et cetera. For illustration of the necessity and benefit from merging ontologies refer to Table 61. It shows the considered data of the respective ontologies and the level of detail thereof.[24]

[21] Source: https://sargon-n5geh.netlify.app/
[22] Stavropoulos, T. G., Vrakas, D., Vlachava, D., & Bassiliades, N. (2012). BOnSAI: A smart building ontology for ambient intelligence. *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics - WIMS '12,* 1. https://doi.org/10.1145/2254129.2254166
[23] Syed Gillani, Frederique Laforest, Gauthier Picard. A Generic Ontology for Prosumer-Oriented Smart Grid. *3rd Workshop on Energy Data Management at 17th International Conference on Extending Database Technology*, Mar 2014, Athènes, Greece. 6p. ⟨emse-00948316⟩
[24]J., Larrinaga, F., & Curry, E. (2017). *A Unified Semantic Ontology for Energy Management Applications*. 12.

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

Data	ThinkHome	SAREF4ENER	BOnSAI	ProSGV3
Infrastructure technical data	H	L	M	M
Energy consumption systems data	H	M	L	H
Energy performance data	H	H	H	H
Sensors/actuators data	H	M	M	M
Energy stakeholders data	M	—	L	L
Weather/climate data	H	L	L	M
Geographical data	—	—	—	L
Environmental data	M	—	M	—
Distributed energy sources data	M	L	L	M
Energy demand-response operations	—	M	—	L

Table 61 - Level of detail of energy domain data (H=high, M=medium, L=low) that the respective ontologies consider.

If it is discovered that some concepts used in well-known standards such as USEF[25] and OpenADR[26] are not covered by such a networked development methodology, emerging ontologies can be integrated after the fact.

One such focal domain is Demand-Response (DR), as is the case with OpenADR standardisation for which now exists a dedicated OpenADR ontology[27]. Another example for an ontology concerning DR flexibility is MIRABEL[28], an ontology for smart grids that aims at leveraging additional flexibility potential by considering additional energy actors such as residential actors.

Digital twins are another hot topic, e.g., for the energy distribution and electrical vehicles (charging stations, parking meters, traffic systems) in smart cities. Notably, Microsoft recently released their smart buildings[29], smart cities[30] and energy grid ontologies[31] for their Azure Digital Twins Platforms. The former is based on RealEstateCore, whereas the latter two are based on a CIM NGSI-LD data model defined under the Smart Data Models initiative and on CIM, respectively.

[25] See also: https://www.usef.energy/
[26] See also: https://www.openadr.org/
[27] Fernandez-Izquierdo, A., Cimmino, A., Patsonakis, C., Tsolakis, A. C., Garcia-Castro, R., Ioannidis, D., & Tzovaras, D. (2020). OpenADR Ontology: Semantic Enrichment of Demand Response Strategies in Smart Grids. *2020 International Conference on Smart Energy Systems and Technologies (SEST)*, 1–6. https://doi.org/10.1109/SEST48500.2020.9203093
[28] Verhoosel, J.P.C., Rothengatter, D.C.F., Rumph, F.J., & Konsman, M.J. (2012). *An ontology for modeling flexibility in smart grid energy management.* Retrieved 16 July 2021, from http://resolver.tudelft.nl/uuid:bfe13c5f-a5db-4056-bc31-bde4f5ddd643
[29] See: https://techcommunity.microsoft.com/t5/internet-of-things/realestatecore-a-smart-building-ontology-for-digital-twins-is/ba-p/1914794
[30] See: https://techcommunity.microsoft.com/t5/internet-of-things/smart-cities-ontology-for-digital-twins/ba-p/2166585
[31] See: https://github.com/Azure/opendigitaltwins-energygrid/

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

4.7. Conclusion

The presented standardizations (ontologies, agile standardizations, etc.) cover, at least initially, the needs of the LSPs. However, it is possible that during the execution of the use cases complementary standardization would be required. Agile standardization could cope with most of these needs when this cannot be covered by existing, open, and adopted standards.

The presented options, smart data models, SAREF4ENERGY, SSN ontology, EFI, SARGON and others, are mostly complementary in several aspects. As work on the BD4NRG architecture continues, a holistic approach will be carried out and the data clusters presented in chapter 3, together with the specific needs of the use cases will help on the final decision of the standards to be adopted.

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

5 Annexes

5.1. List of models from Smart Data Models initiative

The following smart data models describe a part of the domains of interests of the LSPs.

- ACMeasurement.
- AirQualityObserved
- Alert.
- Anomaly.
- Battery.
- BatteryStatus.
- Building.
- BuildingOperation.
- Device.
- DeviceModel.
- ElectroMagneticObserved
- EnergyConsumer
- EOAnalysis
- EODataHub
- EOGeoDataLayer
- EOInstrument
- EOProduct
- EOSatellitePlatform
- EVChargingStation.
- GreenEnergyGenerator.
- GreenEnergyMeasurement
- IndoorEnvironmentObserved
- InverterDevice.
- PhotovoltaicDevice.
- PhotovoltaicMeasurement.
- SmartMeteringObservation.
- StorageBatteryDevice.
- StorageBatteryMeasurement
- Substation.
- TechnicalCabinetDevice.
- ThreePhaseAcMeasurement.
- Vehicle.
- VehicleModel.
- WeatherForecast.
- WeatherObserved.

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

5.2. List of terms defined in SAREF4ENER ontology

This list includes all the terms described in the SAREF4ENER ontology. Potentially they could be reused in the different LSP. Other terms not explicitly related to Energy could be contained in the other members of the SAREF family of ontologies.

- Activation delay
- Activation delay duration description
- Active duration description
- Active duration max
- Active duration min
- Active duration sum max
- Active duration sum min
- Actuator level
- Actuator switch
- Alternatives group
- Anticipation duration description
- Default duration
- Default duration description
- Device
- Duration uncertainty
- Duration uncertainty description
- Earliest start time
- Elapsed slot time
- End time
- End time duration Description
- Energy
- Energy expected
- Energy max
- Energy min
- Energy skewness
- Energy Standard Deviation
- Event action consume
- Event action produce
- Event state consume
- Event state produce
- Latest end time
- Load Control event action
- Load control event data
- Load control event state
- Load control state data
- Max activation delay
- Max anticipation
- Max duration
- Min duration
- Pause duration description
- Pause duration max
- Pause duration min
- Pause time
- Power
- Power expected
- Power max
- Power min
- Power profile
- Power sequence
- Power sequence state
- Power skewness
- Power standard deviation
- Remaining pause time
- Remaining slot time
- Resume cost estimated
- Resume energy estimated
- Slot
- Slot time duration description
- Start time
- Start time duration description
- State
- Time period
- Time period duration description

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

5.3. Data Template

Pilot partners provided an overview of the data assets that will be available in BD4NRG using the template shown in the figure below. Besides the description of the data, partners were asked to provide aspects related to the 4Vs of big data: Volume, Velocity, Variety and Veracity.

BD4NRG Pilot data

This document provides a data template, to be used by pilots' partners, to provide information about the available data sources within the context of its pilot. Moreover, it presents some indicative examples, in order for the input required in each case to be clear. In that regard, pilot partners are kindly asked to provide information regarding the data they possess, in terms of data volume, velocity, variety and veracity, thus ensuring that the data that are about to be used, belong to the big data spectrum. This stage of data description takes place in the context of the preliminary collection of a sample of data, to understand its structure and information, while a second cycle will be implemented.

LSP No: X
BD4NRG Partner: X
Date: XX/XX/XX

Description					Volume						Velocity				Variety					Veracity		
Dataset name	Dataset type	Variable name	Variable description	Measurement Units	GB/year	Starting date	Ending date	Sampling frequency	Percentage of missing values	Storage format	Storage frequency	Read frequency	Update frequency	Bandwidth limitation	Communication interfaces	Interface documentation	Need processing	Data structure	Related standard	Ownership	Reliability	Security of the transmission
XXX	Semi-static																					
XXX	Static																					
XXX																						
XXX																						
XXX																						
XXX																						
XXX																						
XXX																						
XXX																						
XXX																						
XXX																						

1

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

5.4. Overview of Data availability in LSPs

Data Type	Availability in LSPs
Energy Production	LSP 6 • AC Meter readings LSP 7 • AC Meter readings • DC Meter readings • ESS (Energy Storage System) Control data • Load Simulator • PV Simulator LSP8 • Kos-Kalymnos Power Generation System Data LSP9 • Solar Lab PV installation data. • Solar Demonstration Platform data. LSP5 • Open data from Dutch DSOs
Transmission / Distribution System (Energy Grid Assets/Devices)	LSP1 • Transmission tower & overhead lines Right-of-Way LiDAR scan data • Circuit breaker incident data LSP2 • SCADA data • Technical data • OLM data • DGA measurements • Oil quality measurements • Power transformer electrical measurements • Work orders for maintenance LSP4 • SCADA data LSP5 • Enexis MV measurements LSP 6 • Asset Condition Data • Meter Events
Energy Consumption	LSP3 • Charging station data • EV data LSP 6 • Electricity Consumption LSP8 • Water Pumping Stations data • Residential Energy Demand Data

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

Data Type		Availability in LSPs
		LSP9 • Martim Longo School Power Consumption data. LSP10 • Data from management systems: BIM, HEMS, BEMS, DEMS • Invoices LSP11 • BEMS • Invoices LSP12 • Heat energy consumption Invoices • Electricity consumption
Environment & Context	**Building assets / Indoor Environment**	LSP5 • Residential demand response assets LSP10 • Occupant behaviour
	Environmental Information	LSP3 • Current weather data LSP 4 • Weather forecast data • Weather reanalysis data LSP 6 • Weather forecast data LSP7 • Current weather data LSP9 • Solar Lab Weather data (Local Weather Station) LSP10 • Open Weather data from meteorological services. LSP11 • Open Weather data from meteorological services. • Lidar data LSP12 • Temperature, carbon dioxide, humidity
	Administrative Data / Public Services	LSP10 • Cadastre data • EPC registers • Energy DataHub Castilla y León • Building audits LSP11 • Cadastre data • License files • Building EPC

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