A Platform Ecosystem Providing New Data For The Energy Transition

MARKUS DUCHON, JESSY MATAR, MAHSA FARAJI SHOYARI, ALEXANDER PERZYLO, and INGMAR KESSLER, fortiss GmbH, Germany
PATRICK BUCHENBERG, PHILIPP KUHN, and THOMAS HAMACHER, Technische Universität München, Germany
THORSTEN SCHLACHTER and WOLFGANG SÜSS, Karlsruhe Institute of Technology (KIT), Germany
NGUYEN XUAN THINH, HANIYEH EBRAHIMI SALARI, and JASMIN LATKO, Technische Universität Dortmund, Germany
MINSHENG XU, MAXIM SHAMOVICH, DOMINIK SCHLÜTTER, and JÉRÔME FRISCH, RWTH Aachen University, Germany
KUSHAGAR RUSTAGI and MARKUS KRAFT, Computational Modelling Pirmasens GmbH, Germany
CAROLIN AYASSE and FLORIAN STEINKE, Technical University of Darmstadt, Germany
MICHAEL METZGER and LAURA KUPER, Siemens AG, Technology, Germany

There is a great need for high-quality and comprehensive data in the energy sector. This data is collected and preprocessed at considerable expense and is not only required for research, but also by planning offices and other industries in connection with planning activities, such as the creation of municipal heat planning. The NEED ecosystem will accelerate these processes establishing an efficient, robust, and scalable energy data ecosystem. Heterogeneous energy-related data sources will be brought together and automatically linked consistently across different sectors as well as temporal and spatial levels. In this context, existing data sources will not be replaced but rather integrated into the NEED ecosystem as dedicated sources including a semantic description on how to utilize them. In addition to conventional data sources from the various planning levels, we envision a quality assessment scheme based on the FAIR criteria. In reality, we are often faced with missing data, too. To close this gap we explore data-driven, model-driven, AI-based, and tool-driven generation of synthetic data. These heterogeneous data sources will be interlinked using ontology modules which will be represented in a knowledge graph. Via a semantic API, queries will be generated to identify the required data sources, which will be orchestrated to provide the data needed. This will enable researchers, planners, and others including their tools to interact with the NEED ecosystem, while a tool proxy will be able to translate the resulting data into proprietary formats, required by some tools to operate. The NEED ecosystem is planned to be a robust, easy-to-maintain, and flexible infrastructure to enhance planning energy measures at different spatial levels and with different time horizons. We envision to evaluate our NEED approach for the transparent provision of data by integrating relevant data sources as microservices, definition and analysis of application scenarios in the planning domain, as well as the integration of various tools for different planning purposes. With these elements, we will be able to quantify the efficiency of data procurement and demonstrate the functionality of the approach using practical use cases.

Authors' addresses: Markus Duchon, duchon@fortiss.org; Jessy Matar, matar@fortiss.org; Mahsa Faraji Shoyari, farajishoyari@fortiss.org; Alexander Perzylo, perzylo@fortiss.org; Ingmar Kessler, ikessler@fortiss.org, fortiss GmbH, Guerickestr. 25, München, Germany, 80805; Patrick Buchenberg, patrick.buchenberg@tum.de; Philipp Kuhn, pkuhn@tum.de; Thomas Hamacher, thomas.hamacher@tum.de, Technische Universität München, Lichtenbergstraße 4a, Garching b. München, Germany, 85748; Thorsten Schlachter, thorsten.schlachter@kit.edu; Wolfgang Süß, wolfgang.suess@kit.edu, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, Karlsruhe, Germany, 76131; Nguyen Xuan Thinh, nguyen.thinh@tu-dortmund.de; Haniyeh Ebrahimi Salari, haniyeh.ebrahimi@tu-dortmund.de; Jasmin Latko, jasmin.latko@tu-dortmund.de, Technische Universität Dortmund, August-Schmidt-Straße 10, Dortmund, Germany, 44227; Minsheng Xu, xu@e3d.rwth-aachen.de; Maxim Shamovich, shamovich@e3d.rwth-aachen.de; Dominik Schlütter, schluetter@e3d.rwth-aachen.de; Jérôme Frisch, frisch@e3d.rwth-aachen.de, RWTH Aachen University, Mathieustraße 30, Aachen, Germany, 52074; Kushagar Rustagi, krustagi@cmpg.io; Markus Kraft, mkraft@cmpg.io, Computational Modelling Pirmasens GmbH, Delaware Avenue 1-3, Pirmasens, Germany, 66953; Carolin Ayasse, carolin.ayasse@eins.tu-darmstadt.de; Florian Steinke, florian.steinke@eins.tu-darmstadt.de, Technical University of Darmstadt, Landgraf-Georg-Str. 4, Darmstadt, Germany, 64283; Michael Metzger, michael.metzger@siemens.com; Laura Kuper, laura.kuper@siemens.com, Siemens AG, Technology, Otto-Hahn-Ring 6, München, Germany, 81739.

1 INTRODUCTION

The procurement and provision of data is still a very time-consuming and cost-intensive part of planning energy technology systems and is estimated by the project partners to account for 30 % of the total costs. Depending on requirements and use cases, different data—such as LoD2 data[1], information on existing infrastructures, geodata and cadastral data, consumption curves, primary energy requirements, plant data, information on mobility or geothermal potential—have to be procured and preprocessed accordingly for dedicated applications or tools. This process of obtaining and preparing data and information of a sufficient quality and up-to-dateness for the application must be repeated for each and every planning case. Additionally, in order to advance the energy transition, cross-connections must also be taken into account. Consequently, both, the heat and electricity side must be considered in all planning perspectives, e.g., from the perspective of a building or with regard to the infrastructure. This also applies to other energy vectors such as gas, mobility, or water. In many areas, however, hardly any data of sufficient quality is available in machine-readable form and the origin and transparency of data generation and quality are not reliable. With this in mind, the publicly funded NEED research project[2] was launched in September 2023 to develop an energy data ecosystem for future energy planning with numerous research institutions and industrial partners.

With the help of the NEED ecosystem, the planning basis will be made digitally available in the form of data from different levels and domains and linked with each other by the application of ontologies

[1] 3D building models for Germany from Federal Agency for Cartography and Geodesy
[2] https://enargus.de/pub/bscw.cgi/?op=enargus.eps2&q=%2201256602/1%22&v=10&s=13, accessed 10.05.2024

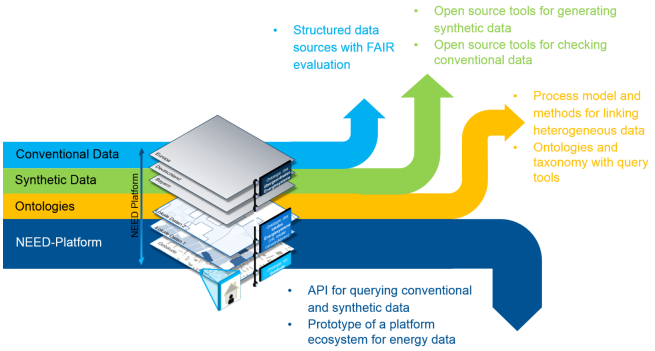

Fig. 1. Multilayered NEED ecosystem: The NEED-platform aims to provide a communication interface for energy data. To provide the interface, underlying ontologies link heterogeneous data sources. The ecosystem will increase accessibility and coherence of data of different aggregation levels, using conventional data sources complemented by synthetic data.

resulting in a knowledge graph. This will create a modular system for end-to-end planning tasks from the building to the infrastructure, which enables automated, model-based analyses across system boundaries. In addition to an improvement in quality through transparent, digital, and verified data, the costs and time required for data acquisition and preparation may be significantly reduced.

In this context, the aim is not to replace existing data sources, but rather to integrate them into the ecosystem as dedicated sources. In addition to conventional data sources (such as energy atlases, state offices, grid operators, building data, geothermal energy deposits, weather data, census data) at the various planning levels (e.g., buildings, districts, regions), we aim to explore ways of closing existing data gaps with synthetic data. By integrating conventional data and deriving synthetic data, the NEED ecosystem may provide a robust, easy-to-maintain, and flexible tool for deriving energy measures at different spatial levels without losing sight of the overall picture.

Finally, the partners' tools and models should access the required data via semantic queries and suitable interfaces in order to fulfill the respective (planning) tasks. This illustrates the NEED approach to the transparent provision of current data, particularly in the examples of heat management planning and the creation of a dynamic energy usage plan. The following subgoals are pursued:

- Subgoal 1: Supplement and integrate existing data sources and data formats, such as the energy atlas of Bavaria[3] or the energy atlas of Thuringia[4], LoD2 data of Bavaria[5], etc., to integrate different conventional data sources.
- Subgoal 2: Provide and validate synthetic data created through derivation and aggregation.
- Subgoal 3: Bringing together different sources and levels on the topic of energy in a knowledge graph to realize semantic queries and, based on this, to set up monitoring, if possible in real time, and to run knowledge-augmented 'what-if scenarios'.
- Subgoal 4: Validation and verification by means of practical applications for energy planning based on conventional and synthetic data using real examples in order to demonstrate the benefits and applicability with the aim of stabilization.

The paper is structured as follows. Section 2 will introduce the stakeholders addressed in our work including the main application scenarios. Section 3 deals with conventional data sources and the corresponding investigations. In Section 4, synthetic data for closing existing gaps in the conventional data or to validate the developed data-driven approaches is introduced. To interconnect the available and existing data sources with actual planning tools, in Section 5, we present the concept of ontology models resulting in a knowledge graph to query relevant information. To combine the so far presented building blocks we propose a distributed service-based architecture in Section 6 before we summarize the work conducted and give a brief outlook.

2 USE CASES AND CASE STUDIES

In order to develop the NEED ecosystem addressing actual challenges and problems of practice, use cases and case studies are defined. The use cases describe what methods, tools, and data are used, as well as the respective customers addressed. One or two specific questions for each use case are elaborated, which are then used as guidance. Based on the questions, the necessary analysis methods, tools, and data are identified.

The description of used methods and tools provides information about the interfaces to the NEED ecosystem. That also includes the data requirements for providing information (e.g., population development, spatial development, time series for forecasts, geographical allocation, defined units). As a result, the overall requirements regarding data and interfaces to the NEED ecosystem can be specified. Within the research project, the partners agreed on focusing on the following customers:

- Homeowners, housing associations, energy consultants
- SMEs (small and medium-sized enterprises)
- Network operators
- Municipalities, communities

The first two types of users reflect the point of view from an individual perspective either at residential or industrial level. The third and fourth types of users represent the supply side and are responsible for operating energy networks and ensuring energy supply.

The scientific partners and especially the industry partners have a variety of tools available for this purpose. With regard to the different tools, the resulting use cases are summarized in application scenarios, which represent a thematic cluster. The following four application scenarios were developed by the project partners to take energy planning requirements into account:

- Energy utilization and municipal heat planning
- Network planning (electricity and heat)
- District planning
- Energy consulting (renovation/transformation/electrification)

[3]https://energieatlas.bayern.de/, accessed on 10.05.2024
[4]https://karte.energieatlas-thueringen.de/, accessed on 10.05.2024
[5]https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=lod2, accessed on 10.05.2024

The creation of these four application scenarios covers both the expertise of the consortium partners and the demand of the above-mentioned customers. In addition, we have ensured that a set of suitable case studies exist for each of the application scenarios to validate the benefits of the proposed NEED ecosystem over the course of the project.

2.1 Case Studies

A case study is defined as a real project carried out by the consortium partners in collaboration with local stakeholders. The case studies are intended to demonstrate and validate the benefits of the NEED ecosystem at a later stage of the project. Each case study is assigned to a previously defined use case with an associated customer. There is at least one case study for each use case.

Various industrial partners in the NEED consortium are already working on projects that can be assigned to a specific application scenario. Some of these projects will be converted into case studies and then reworked with the help of the NEED ecosystem. The aim is to show how the planning process for the various use cases changes through the use of the NEED ecosystem. The aim is to significantly reduce the time and effort required to record and preprocess the input data, as the relevant information may be efficiently retrieved in the desired spatial and temporal resolution and in the desired format via suitable NEED interfaces and proxies.

In addition to relying on projects already being worked on by industry partners, new methods for processing the planning tasks of the various use cases are also being investigated. This also includes investigations into new efficiency potentials made possible by the NEED ecosystem. The possibility of querying data in different temporal and spatial resolutions should facilitate the inclusion of synergy effects of neighboring units in a variety of planning tasks. In addition, research is being carried out into the automated provision of data, e.g., for energy consulting for residential buildings or for municipal heat planning.

A selection of case studies is presented below:

- Municipal heat planning for a city in southern Hesse
- Detailed heat grid planning in a city in southern Hesse and in a city in Bavaria
- Generation of synthetic electricity distribution networks in a city in northern Bavaria
- Automated data provision for energy efficiency consulting of (residential) buildings

Within these case studies, we aim to increase the level of automation in the planning processes. For example, for heat grid planning, we intend to develop a method to handle the (computational) complexity of topology design to make automated heat grid planning applicable in real planning processes. However, the executions of the case studies and the integration of automated processes rely on a wide range of different data sources at the various planning levels. Some data are publicly available, others are provided by government institutions, others are not available at all. Nevertheless, it is no trivial task to collect and preprocess the right data in the right quality and up-to-dateness. In this context, the project aims to analyze and process existing, so-called conventional data sources to integrate them into the NEED ecosystem.

3 CONVENTIONAL DATA SOURCES

In the context of the NEED project, conventional data refers to data that is already accessible and typically published by the authors of the data in different data sources, e.g., web-based map services such as the *Energieatlanten* (energy atlases) or online registries such as the *Markstammdatenregister* (master data on the electricity and gas market)[6]. Usually, these data sources serve specific purposes, resulting in limited connections between sources. Even though in some instances the data of different sources overlap (for example some of the energy atlases include data about global sun radiation which is a dataset released by the *Deutscher Wetterdienst*, Germany's institution for meteorological services[7]), sources like energy atlases only include data for a specific time frame or a specific spatial extent instead of holistic datasets.

Given the broad spectrum of data sources with ready-to-use data, it is necessary to gain an overview of the data in order to integrate them into the NEED ecosystem. We are therefore working on finding a uniform characterization for all data in order to assess data quality and characterize further criteria in the variety of data sources and authors of the data. Testing our characterization approach requires the use of a range of data sources with comparable datasets, which led to the decision to use the Energy Atlases.

3.1 Evaluation of Energy Atlases

An energy atlas typically refers to a web-based map service hosted by a federal state of Germany. Generally, the hosted data relates to the field of energy planning in the broad sense and ranges from simple geometric data used to describe the visualized area to the global sun radiation over the course of one year. As of now, not every state has an energy atlas and those that do, all have their own instead of one energy atlas for the whole country (see Figure 2). The atlases differ regarding the hosted content and data quality. Since all these atlases were probably created with similar intentions, working with them would be much more efficient if they all followed the same standards. In order to realize a common standard, it is necessary to gain an understanding of what data is contained in the individual energy atlases and in what quality. To this end, the energy atlases are analysed and two matrices are introduced to characterize and evaluate the individual data sets and their metadata. The first step is to develop the columns of the two matrices. The data to be analyzed in the matrices consists mainly of geodata. While most of the data comes from the field of energy planning, there is a focus on the inclusion of basic geodata, as this represents an essential basis in any planning process. The matrices must therefore be designed in such a way that they allow data from different areas to be analyzed. This led to the columns being strongly oriented towards the way in which data is analyzed in the geoinformation sciences. To ensure the reusability of the data and to allow users to easily assess the quality of the data, a strong focus is placed on the spatial and temporal aspects of the datasets and their metadata.

[6] https://www.marktstammdatenregister.de/MaStR, accessed on 10.05.2024
[7] https://www.dwd.de/DE/leistungen/solarenergie/solarenergie.html;jsessionid=2A01EAA2CFC9627F02AC72EF3D5DDDD9.live21074?nn=16102, accessed on 29.04.2024

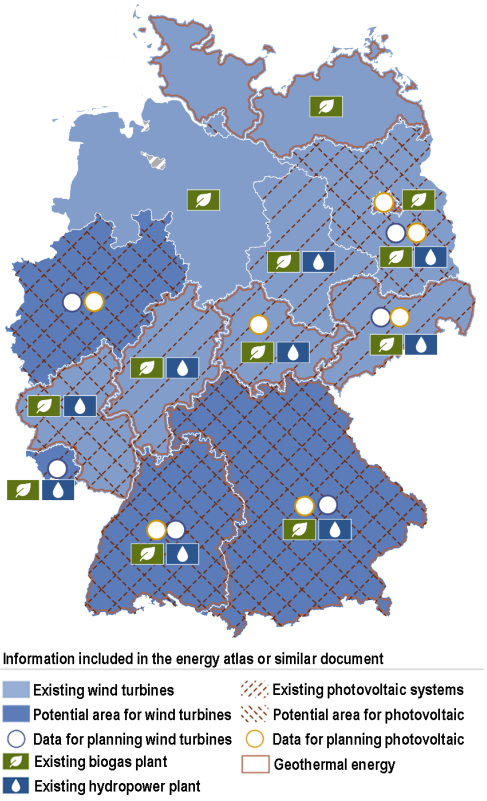

Fig. 2. Information included in the energy atlases of Germany's federal states.

The first matrix focuses on the metadata to gain a basic understanding of the corresponding energy atlas and its quality. It includes the following aspects:

- The name of the dataset as it is shown in the energy atlas,
- Contact information of the dataset's author,
- Information about when the metadata was updated last,
- The data's usage license,
- The hyperlink to the metadata, and
- The option to evaluate the data regarding the FAIR principles with certain indicators for each principle.

The above-mentioned 'FAIR principles' evaluate the findability, accessibility, interoperability and reusability of data and serve as a benchmark for data quality and applicability [Wilkinson et al. 2016]. For the NEED ecosystem, we pre-select FAIR principles that are relevant in the context of energy planning processes and evaluate them for the datasets under consideration according to [Bahim et al. 2020]. We plan to make the evaluation of these principles together with the datasets available to the users of the platform in order to enable an individual evaluation of the relevance of the individual principles in specific planning processes.
The second matrix focuses on the specific datasets and their contents, and the following aspects are included in the matrix:

- The name of the dataset as it is shown in the energy atlas,
- The topic the dataset that it can be associated with, such as electricity, mobility, or heat energy,
- Information about the dataset's source, e.g., if it is commercial data or open data,
- The classification in spatial data with or without references to a specific subject or research field,
- The geometric dimension and whether the dataset includes temporal information,
- The specific content of the data such as buildings, different kinds of infrastructure, etc.,
- The data's spatial extent, ranging from single georeferenced locations to a whole country,
- The data's resolution or scale,
- Information about how often the data is updated,
- Information about when the dataset was initially released,
- Information about when the data was updated last,
- Information about the data format,
- Information whether the data can be downloaded,
- Contact information of the dataset's author,
- A hyperlink to the dataset in the energy atlas, and
- The option to evaluate the data regarding the FAIR principles with certain indicators for each principle.

Each of these aspects serves as one of the columns of the matrix and all information must be added to the matrix for each relevant dataset. Once all information on the datasets of all energy atlases has been collected in the matrix, it is possible to compare the different atlases. The comparison can be used to make a proposal to the institutions responsible for the atlases in order to identify possible weaknesses and deficiencies and to establish a nationwide standard. In addition, the matrix can also be used to characterize data from other data sources. Since all data is characterized in the same way, it can be more easily implemented into the NEED ecosystem for future use without the need to implement different parsers for each data source. The data itself is not necessarily included in the same way as it is usually retrieved, e.g. on an energy atlas website. Instead, the data can be mapped to specific areas based on the geographical information about the datasets. When information about these areas is retrieved, the datasets and their respective entries from the matrices can be displayed and accessed via hyperlinks or other references to where they were originally published. In this way, additional datasets can be added to the ecosystem, even if not all data can be downloaded or accessed in the form of a Web Map Service. In the following, we describe the *Marktstammdatenregister (MaStR)* as another example of conventional data.

3.2 Open-MaStR

Germany's publicly available *Marktstammdatenregister* is a register that keeps track of energy units (including power and gas). It is provided by the German Federal Network Agency *(Bundesnetzagentur / BNetzA)* and is updated on a daily basis. The *MaStR* open dataset can be browsed online on the website of the *BNetzA*. To facilitate access to the database, a Python package called *open-mastr* has been developed by the *RLI* and *fortiss* to provide an interface to improve

the usability of the register. The package includes methods to clean and write the data into a local database[8].

The package provides a Python interface for accessing data via the bulk download and the web service API, and methods to clean the data. Through the bulk method, one can download all units (e.g., all solar farms in Germany) and through the API method, one can retrieve specific information regarding single units (requires registration for an API token). The cleaned data is then written into a database. The *open-mastr* package hence provides easy access to the dataset, which is especially useful in the community of energy system researchers.

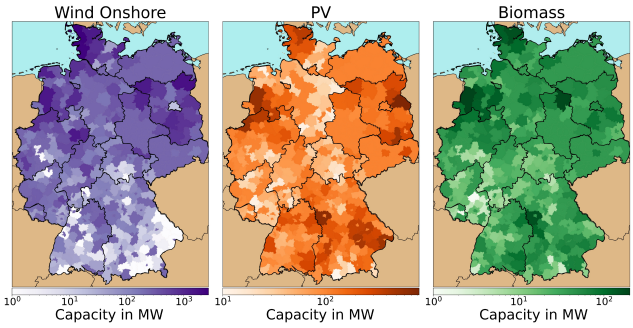

Fig. 3. Installed capacity in Germany per district in 2024 extracted from the *Marktstammdatenregister* using *open-mastr*.

Easy and direct analysis of various energy sources installed nationwide, in every state, district, and municipality can be tracked through this package. Figure 3 illustrates the capacity provided from wind, PV, and biomass in 2024 in Germany per district. The data found in the *Marktstammdatenregister* are manually entered, resulting in a large potential of errors and data gaps. The latter issue will be tackled in the next section, where we introduce synthetic data generation and its importance in filling the gaps of conventional data sources.

4 SYNTHETIC DATA

Despite the considerable amount of conventional data presented in Section 3, there are also significant data gaps. In some cases, needed datasets are not available at all, in other cases the necessary temporal or spatial resolution is lacking, or existing datasets cannot be used or have to be modified for data protection reasons, for example.

In addition, datasets from different sources often have different technical and semantic representations, which is why it is necessary to map, link, and harmonize datasets. It therefore makes sense not only to think about generating *data*, but also to create, provide, and use *metadata* for all (both conventional and synthetic) datasets to ensure data can be (automatically) converted, mapped, linked, etc.

Metadata for both synthetic datasets and processed conventional datasets should contain traceability of their origin and all processing steps applied (so-called *provenance data*). This is essential for the reproducibility and transparency of the models, simulations, and plans created along the data.

4.1 Strategy for the Creation and Use of Synthetic Data

The lack of reliable data and the need for preserving the privacy of the data are major concerns, e.g., when working with low-voltage networks or residential data, particularly when such data is unavailable to the general public or cannot be measured, recorded, or documented. However, reliable outputs can be achieved and gaps can be filled with synthetic data created using the characteristics of conventional data.

Our primary goal is to create an ecosystem that provides and integrates conventional and synthetic data necessary for energy planning. In this section, we focus on exploring the current ideas related to synthetic data generation, with a particular focus on exploring various methodologies. We introduce methods such as image processing, data fusion, and AI, ensuring validation and verification of the generated data [Dankar and Ibrahim 2021].

The project partners will focus on three strategies in particular:

- Disaggregation of data that is available at a lower temporal or spatial resolution
- Reconstruction of data using established planning principles
- The utilization of unconventional data sources, such as aerial and satellite photos

First and foremost, we examine methodologies used in diverse fields for synthetic data creation. We explore methods from geoinformatics to be implemented for energy planning purposes. We identify gaps and challenges in current conventional data sources in the energy sector, and use synthetic data generation approaches to fill these gaps.

Our proposed methodologies for generating synthetic data encompass a range of approaches. Starting from the use of image data to extract information and characteristics relevant to energy systems. Additionally, we tackle methodologies that harness data sources to derive information, particularly in the context of building sectors and regional analysis, through aggregation and fusion techniques.

4.2 Methods for the Generation of Synthetic Data

We will also explore both the standard methods, such as Random Oversampling (ROS), Cluster-Based Oversampling, and Gaussian Mixture Models, as well as deep learning methods such as Generative Adversarial Networks (GAN) used for generating synthetic data (see Figure 4). Deep learning has led to the emergence of several promising techniques for creating synthetic data, most notably the GAN. Furthermore, GAN can produce new synthetic samples that closely resemble the original dataset's underlying data distribution. Since the time it was proposed, the architecture of GAN has been modified based on the use case or field of application. Hence, we will delve into finding the most suitable architecture for generating synthetic data.

4.2.1 *Random Oversampling (ROS).*

Oversampling is a method in which the minority classes are duplicated. ROS is one of the classical approaches to oversampling,

[8]https://github.com/OpenEnergyPlatform/open-MaStR/, accessed on 10.05.2024

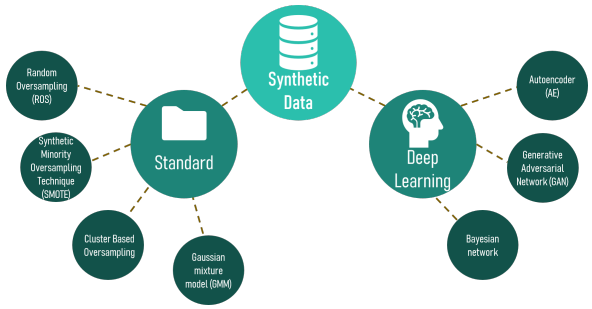

Fig. 4. Different methods for generating synthetic data [Figueira and Vaz 2022].

which expands the dataset with new observations by randomly selecting a replacement sample from the minority class. Although this is the most direct approach to growing a dataset, this method merely replicates the existing samples rather than generating new ones [Batuwita and Palade 2010].

4.2.2 Synthetic Minority Oversampling Technique (SMOTE).
SMOTE is an oversampling technique in which new instances are generated for each training observation by selecting points at random. There are two components to the SMOTE algorithm: the selection mechanism and data generation mechanism [Chawla et al. 2002]. In the selection mechanism, a minority class observation and its nearest neighbors are selected at random. The data generation mechanism is responsible for generating synthetic data. Despite its advantages, SMOTE has certain limitations, such as generating noisy data. Hence, there are many variants to the classical SMOTE method that address this issue, such as the Borderline-SMOTE, Safe-level SMOTE, and Adaptive Synthetic Sampling approach (ADASYN).

4.2.3 Cluster-Based Oversampling.
Cluster-Based Oversampling, first proposed by [Jo and Japkowicz 2004], involves clustering the training data in the minority and majority classes separately and then applying ROS to each cluster. This method aims to improve the between-class and within-class imbalances.

4.2.4 Gaussian Mixture Model (GMM).
In cases where the dataset displays multiple areas of elevated density, a single Gaussian model may struggle to adequately fit the data. This is where GMM proves to be particularly useful. GMM is a probabilistic model that assumes that the data is a mixture of many Gaussian distributions, representing different subpopulations within a dataset, each of which contributes a certain weight to the whole distribution. The model learns these subpopulations through training using the Expectation Maximization algorithm.

4.2.5 Bayesian Networks.
A Bayesian Network, also known as a *belief network*, is a type of graphical model that represents the joint probability distribution for a group of variables [Young et al. 2009]. The two main components of the Bayesian model are a set of graphical structures and a set of conditional probability distributions. The graphical structure is a set of nodes. Each node in a Bayesian Network corresponds to a probability distribution that quantifies the likelihood of a variable taking on different values given the values of its parent variables [Heckerman et al. 1995].

4.2.6 Autoencoders (AE).
An autoencoder is a type of artificial neural network used in unsupervised learning. It consists of an encoder network and a decoder network that work together to learn the diverse representation of the input dataset [Ackley et al. 1985]. The encoder compresses the input data into low-dimensional data, called 'latent space', and the decoder then reconstructs the original input data from the low-dimensional representation. Despite their advantages, autoencoders also suffer from some disadvantages. They may struggle to capture the full diversity of the input data and as a result the generated samples might lack diversity or fail to represent all possible instances in the dataset. When trained on a limited dataset, autoencoders could also suffer from overfitting. This can lead to samples being generated that are very similar to the training data, but do not accurately capture the underlying distribution [Figueira and Vaz 2022].

4.2.7 Generative Adversarial Networks (GAN).
The idea behind GAN is to train two neural networks, a generator (G) and a discriminator (D), simultaneously (see Figure 5). The generator model tries to imitate the underlying original distribution of the data while the discriminator model tries to classify a given observation as real (coming from the original dataset) or fake (generated by the generator model) [Goodfellow et al. 2020]. During the training, the two models compete with each other. Based on the discriminator's feedback, also known as the *loss function*, the generator attempts to improve its performance by generating more realistic samples (see Figure 5). Through this process, both the generator and discriminator models improve their performance, leading to the generation of high-quality synthetic data.

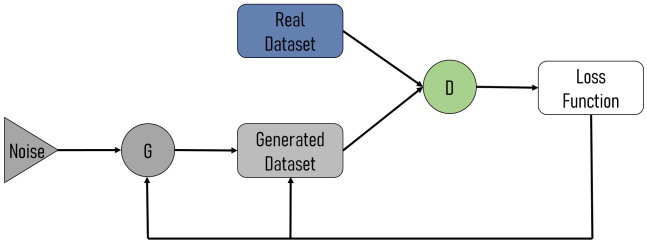

Fig. 5. Overview of GAN. Given a random noise, G generates a set of datapoints, the generated dataset. The generated dataset and the real dataset are fed to D, which labels the output as a loss function. This is then fed back to G and D to improve their performances [Figueira and Vaz 2022].

4.3 Validation Metrics

To ensure the proper validation of synthetic data , we will consider evaluation metrics tailored to our specific use case. Existing metrics often fall short in capturing complex relationships and dependencies within data, such as temporal correlations in time-series data. We will address these limitations essential to ensure a more accurate and rigorous validation process, ultimately leading to higher quality and more reliable synthetic data.

Relying on the existing validation techniques, we will identify key characteristics in conventional data, and design and test new metrics through simulations and empirical analysis. We will employ a combination of statistical tests, visual inspection, and domain-specific criteria for comparison, such as Kolmogorov-Smirnov tests for distributional comparison and Pearson correlation coefficients for relationship assessment [Whitnall et al. 2011]. By providing a comprehensive assessment, these new metrics will enhance the reliability and applicability of synthetic data, ensuring it meets the rigorous standards required for its intended use.

4.4 Ontology-Based Descriptions for Spatio-Temporal Compliance

Ontology descriptions can guide synthetic data generation to produce data that complies with specific scales. For instance, if the ontology specifies that the data should have a specific resolution, the synthetic data generation algorithm can be configured to produce images at this resolution. This ensures consistency with real satellite imagery, facilitating seamless integration into existing datasets. In our initial work, we investigate high resolution aerial images and lower resolution satellite images from Sentinel-2 [Spoto et al. 2012] and use these immense monitoring tools to segment solar installations in Germany. Super-resolution images can be synthesized from the widely available Sentinel-2 data, allowing a regular automated update of the MaStR registry inputs.

By leveraging semantically rich annotations, we can generate high-quality synthetic satellite imagery that aligns with specific requirements. The generated data will enhance the robustness and accuracy of machine learning models, fill the data gaps, support comprehensive analysis, and provide valuable insights across various applications, from environmental monitoring to urban planning enriching further the NEED platform.

5 KNOWLEDGE GRAPH

5.1 Challenges of Data Integration in Energy System Planning

Energy system planning is a multifaceted approach that requires a holistic understanding of the systems and processes involved. Numerous researchers have contributed to this field by proposing energy system models and tools aimed at optimizing energy system planning [Eveloy and Ahmed 2022; Henke et al. 2022; Metzger et al. 2021; Prina et al. 2018]. The accuracy and effectiveness of energy system modeling and planning depends on the quantity and quality of input data [Keirstead et al. 2012].

In the context of energy system planning, data availability stands out as one of the fundamental challenges. These input data originate from diverse databases, encompassing both conventional data discussed in Section 3 and synthetic data explored in Section 4. These data span multiple energy domains, including electricity, heating, and cooling, and cover a wide range of technological, economic, social, and political aspects. Cross-regional cooperation, renewable energy source integration, and environmental impact considerations become more prevalent. Thus, the complexity of energy system planning escalates, accompanied by a dramatic increase in data requirements.

Moreover, data quality assumes critical significance in energy system planning. Errors or inaccuracies within data sources can lead to additional efforts, necessitating the use of proxy data or calculation adjustments [Keirstead et al. 2012]. Such discrepancies may ultimately result in flawed models and suboptimal planning outcomes. In addition, ensuring that data aligns precisely with the intended purpose and context is a foundational principle. Valid data faithfully represent the phenomena under study and significantly contribute to effective planning efforts.

Consequently, there is an urgent need for data integration to enhance interoperability and ensure the availability of reliable data.

5.2 Ontology-Based Data Integration

In recent years, Semantic Web Technologies (SWT) and ontology-based data integration (OBDI) have garnered increasing attention as innovative and effective approaches for addressing challenges related to data management in the energy domain [Sicilia and Costa 2017]. OBDI, as an information management system, comprises three key components: an ontology, a set of data sources, and the mapping between them [Liu and Özsu 2018]. The ontology serves as a set of formal, explicit specifications of a common conceptualization that captures a shared understanding of the matters involved [Lork et al. 2019]. It maps the required terminological aspects within a knowledge representation.

In the context of the OBDI, ontologies provide a high-level, conceptual view of the set of data sources [Calvanese et al. 2011]. By constructing a Knowledge Graph (KG) based on ontologies, this framework enables the interlinking of information across diverse domains at data level.

In the NEED project, most data sources initially provide only tabular rows of information. Classification is necessary to integrate these datasets into a larger framework. This will be achieved using a taxonomy constructed from both existing and synthetic data examples. This taxonomy will delineate key concepts within each domain of the NEED ecosystem, facilitating the clustering and extension of these concepts. Additionally, it will enable data mapping into our knowledge network without necessitating direct replication into semantic triples.

The energy and buildings sectors already boast a substantial collection of existing ontologies [Pritoni et al. 2021]. Therefore, the goal of the NEED project is not to create a new standalone ontology, but rather to utilize these existing resources by aligning with and incorporating new concepts as needed.

Thus, the design of the NEED ontology will be modular. The core module will amalgamate shared concepts and terms that are fundamental across all domains considered, as well as recurring attributes and relations among entities. These shared elements will be identified during the initial data acquisition phase, which will be followed by the dissemination and definition of key entities and attributes.

Each domain within the project, such as buildings, electricity, or heating, will have its own module enriched with specialized classes and categories. These will be linked to the broader concepts established in the core module. The configuration of these domain-specific modules will be shaped by the systems they represent and

the available data types, such as georeferenced data, potentially requiring the integration of a common ontology module across different system categories. Furthermore, these modules will be designed to align with other domain-specific ontologies, enabling module-to-module alignments without necessitating full-scale alignment across the entire NEED ontology, which is known in the literature as a hybrid approach.

The final design of the ontology will encompass all necessary data sources and metrics, accurately reflecting the systems under consideration. The modular setup will ensure that the ontology stack can be flexibly instantiated based on the specific needs of each use case and application, thus ensuring adaptability and relevance across a variety of scenarios.

5.3 Implementation of OBDI via Knowledge Graph

After the ontology design phase, our focus will shift to ontology-based data integration. This will involve mapping data sources to the corresponding modules within the domain ontology. The culmination of this process will be the creation of a comprehensive knowledge graph. This graph will play a vital role in supporting tools for planning and analysis within the application layer, enabling them to perform queries and interact efficiently with the data. Such functionality is crucial as it acts as a semantic bridge, seamlessly linking diverse data sources and enabling the application's features without requiring users to know details of the underlying database structures. A diagram of the envisioned framework is shown in Figure 6.

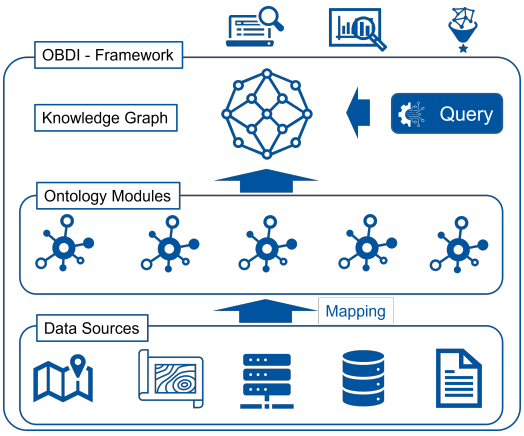

Fig. 6. Possible structure of OBDI in the NEED project.

Implementation of knowledge graph in NEED can be done using The World Avatar (TWA) Project. The World Avatar (TWA) project uses knowledge models and technologies from the Semantic Web to seamlessly integrate data and computational agents [Akroyd et al. 2021]. It offers a general and scalable way to connect heterogeneous data sources to provide an aligned world view, provide up-to-date insights, analyse complex what-if-scenarios, and provide robust control functionalities as a bridge between the physical and the digital worlds, supporting more effective and coordinated decision-making processes.

What sets TWA apart from other approaches is its dynamic behaviour. Computational agents act as autonomous knowledge components to manage and update data. Inputs and outputs from the agents can be semantically annotated to form chains of dependent information. A provenance framework ensures that the consequences of any changes to data are cascaded throughout TWA [Bai et al. 2024a]. This automation, together with input agents that continuously assimilate data feeds into the system, allows TWA to remain current and responsive to new information and scenarios. TWA is scalable by design. The computational agents can wrap around existing software, and new ontologies can be incorporated continuously while maintaining connections to everything existing in real world. The underlying knowledge models (i.e., ontologies) facilitate interoperability and knowledge retention by explicitly codifying domain expertise.

TWA has been applied in a wide variety of contexts including representing molecular scale information to automate calculations [Farazi et al. 2020] [Pascazio et al. 2023], chemical experiments [Bai et al. 2024b] and materials discovery [Kondinski et al. 2022], the development of natural language methods to query data and trigger calculations [Zhou et al. 2021] [Zhou et al. 2022] [Tran et al. 2024], the integration of data from Geographical Information Systems (GIS), Building Information Models (BIM), Building Management System (BMS) [Quek et al. 2024], digitalisation of city planning processes in Singapore [Chadzynski et al. 2021] [Silvennoinen et al. 2023], optimisation of district heating operations in Germany [Hofmeister et al. 2024c], and evaluation of cascading risk to improve the climate resilience of connected infrastructure networks in the UK [Hofmeister et al. 2024a] [Hofmeister et al. 2024b], including the Climate Resilience Demonstrator.[9] TWA separates data and knowledge representation from technical implementation, allowing it to exceed the capabilities of connected digital twins. It is platform-agnostic and open-source, eliminating the risk of vendor lock-in and fostering collaborative and transparent development. It has a distributed architecture that supports safe data access, including local hosting and access control, ensuring data security and privacy. TWA is designed for agents to wrap around existing software, facilitating seamless integration of both current and future technologies. The ideas developed in TWA project will support the development of the underlying knowledge models and technological aspects of the NEED project.

5.4 Semantic Georegistration of Data

As most of the data that are relevant to NEED cover aspects of specific geographic regions or are obtained at specific locations, we target to semantically annotate the data with their geographic references. Ontologies can facilitate the creation of synthetic geographical data by providing a structured and semantically rich framework that defines the entities, relationships, and constraints specific to geographical domains. For this, a hierarchical approach will be established that combines the various relevant scopes for our planning domains from building layouts, to city districts, city limits, administrative districts, states, and the overall country. A

[9] https://digitaltwinhub.co.uk/climate-resilience-demonstrator-credo/, accessed on 10.05.2024

related approach using an OntoCityGML ontology that describes a CityGML-based conceptual schema was used to create a semantic model of the city of Berlin [Chadzynski et al. 2022].

A suitable approach is to rely on the Open Geospatial Consortium's GeoSPARQL 1.1 standard[10], which on the one hand provides a standardized vocabulary for describing geospatial linked data and their corresponding geometric properties and on the other hand a set of extensions to the SPARQL query language to properly interpret the represented models. By employing this technique, represented data in the NEED platform can be easily segmented along the defined geospatial layers, as geospatial properties can be used in the formulation of semantic queries for filtering the data. This will enable the extraction of required data and the structured comparison of different regions.

6 SOFTWARE ARCHITECTURE

In order to supply the application scenarios with the data sources mentioned in the previous sections and link them to a knowledge graph using ontologies, a suitable infrastructure must be created. This is intended to ensure flexible and efficient processing within the framework of a platform ecosystem and requires a modular, scalable, and robust software architecture. Since no existing sources are to be replaced and we are therefore dealing with a highly distributed system, we do not rely on a monolithic architecture but rather on a service-based architecture (SBA) using cloud computing technologies. In this way, we eliminate vendor lock-in and interoperability issues. First, we will highlight the basic concept, and then explain each component of the architecture, before we conclude with a possible instantiation.

Service-Based Architecture (SBA): . SBA is an approach for developing an application as small services, so-called microservices. The characteristics of SBA can be summarized as follows [Lyu et al. 2020; Söylemez et al. 2024]:

- Modularity: The application is broken down into multiple microservices. Each microservices can be developed, deployed, and managed independently. Each microservice is associated with a specific business capability within a certain context boundary, resulting in low complexity and small size.
- Lightweight communication mechanism: The microservices communicate with lightweight mechanisms, often an HTTP resource API.
- Decentralized governance: The scalability of a microservice-based application can be significantly improved through decentralized service governance and data management.
- Agility: SBA makes it easier to adapt to new requirements and change management. In a monolithic architecture, any changes require the whole system to be rebuilt and completely redeployed. However, with SBA, only the affected services need to be rebuilt and deployed independently. Microservices are highly maintainable and testable, making them a great choice for modern software development.

This will enable conventional and synthetic data sources to instantiate their own microservices. These microservices will enable both integration with the ontologies through corresponding API descriptions and the actual, automated retrieval of the requested information. The latter can be understood as wrappers that will offer a NEED-compliant API and translate it into the often proprietary interfaces or implement access to the raw data.

Container Technology: Containerization is a popular virtualization approach where isolation of applications and services occurs at the host level through containers. This provides several advantages for users [Behravesh et al. 2019; Lyu et al. 2020]:

- Lightweight and efficient deployment: Deploying applications using lightweight container images offers practical advantages since these images contain all the necessary dependencies for the seamless operation of an application.
- Portability: A container provides a self-sufficient and executable computational space, encompassing the code, runtimes, system tools, libraries, and configurations necessary for an application.
- Scalability: Containers facilitate agile scaling of applications in response to demand, ensuring optimal resource utilization. This scalability empowers industrial organizations to replicate and distribute cloud-native application instances across various edge devices or servers. This distributed approach enables workload balancing and efficient device utilization, preventing specific infrastructure components from becoming overloaded.
- Secure and reliable: Container platforms improve fault tolerance through robust mechanisms that isolate and manage application failures effectively. In the event of a container failure, it can be quickly restarted or substituted without disrupting the functionality of other containers or the entirety of the infrastructure.

Having a robust and flexible infrastructure is a result of the above characteristics of SBA and container technology. The robustness of a system is an important life-cycle attribute, since it is a strategic attribute that supports business, development, and operation needs. As a result, robust infrastructure uses some techniques in order to detect faults (such as monitoring, heartbeat, condition monitoring, voting, etc.) and recover from faults (such as redundant spares, rollbacks, restarts, etc.), as well as to prevent faults (such as predictive models) [Kazman et al. 2022; Paparistodimou et al. 2020].

A containerized infrastructure also means that NEED components may be deployed relatively easily on other systems and therefore will not necessarily have to be run on a central platform. Individual process chains may be set up through suitable orchestration of the data sources and the integration of preprocessing steps. A simple preprocessing step could be, for example, changing the resolution of time series data. The individual process chain, which will be made up of containers of the data sources in combination with basic functions, may then run locally on the respective client side and thus enable the scaling of the overall system. Figure 7 highlights the proposed NEED architecture, designed to address the shortcomings present in a monolithic architecture. The figure highlights several innovative features as outlined below.

[10] https://docs.ogc.org/is/22-047r1/22-047r1.html, accessed on 10.05.2024

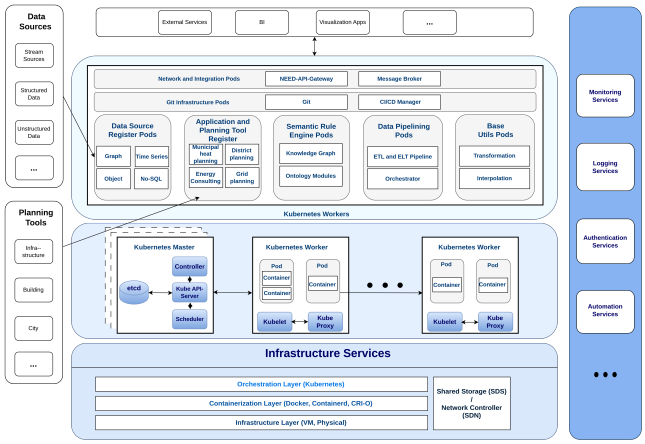

Fig. 7. NEED architecture.

Containerized Deployment Mode: In practice, the proposed NEED architecture will be implemented within a cluster administered by Kubernetes (K8s). This setup will manage various aspects like deployment, maintenance, and scalability of containers, thereby facilitating the management of microservice-oriented applications. Several methods exist for establishing a Kubernetes cluster, with one cost-effective approach being the design and deployment of Kubernetes nodes on predefined cloud infrastructure rather than physical hardware resources. This deployment model typically utilizes Infrastructure as a Service (IaaS). Moreover, to ensure continuous availability of applications, strategies will be employed to enhance the resilience and reliability of the Kubernetes cluster, as depicted in Figure 7. These strategies may include employing multiple nodes for data replication, implementing shared storage solutions, and so forth.

Containerized Microservices as Essential Components: Every NEED application will be housed within its own container, ensuring that its operating environment is isolated from the host system. Consequently, any errors within a container instance do not disrupt the normal functioning of other services. Additionally, container instances can be generated, updated, and terminated with ease and speed, facilitating independent scaling of each microservice as required.

- Data Source Register: Here, the data source microservices will register themselves, so that the platform services will be aware of their data availability and what kind of data (see Sections 3 and 4) will be provided.
- Application and Planning Tool Register: Similar to the Data Source Register, the planning tools or their wrappers that will enable interaction with the Semantic Rule Engine Pods or the knowledge graph will register here.
- Semantic Rule Engine: These pods will hold the required services and functionality for the ontology-based data integration as described in Section 5.2.
- Data Pipelining: The functionality of these pods will enable the orchestration and merging of different data sources, whereby transformations and preprocessing steps may also be inserted in substeps depending on the requirements of the requesting entity.
- Base Utils: They will provide the necessary preprocessing steps, transformations, and manipulations of the data in order to realize seamless orchestration.
- Git and CI/CD Infrastructure Pods: These pods will enable the continuous updating of client and server code throughout the development process. By leveraging these tools, development teams may maintain a streamlined workflow for code integration, testing, and deployment, promoting agility and efficiency in software development cycles.
- Network and Integration Pods: In a microservices architecture, client applications typically interact with various microservices to access different functionalities. However, direct consumption of these microservices necessitates managing multiple calls to their endpoints, which can pose challenges as an application evolves by introducing new microservices or updating existing ones. Moreover, developers may employ different technology stacks and communication protocols for each microservice within their application. Additionally, the elastic nature of the cloud enables services to horizontally scale in response to fluctuating demand, enhancing application resilience. Nonetheless, this scalability necessitates load balancing, simplified service discovery, and features like timeouts and retries for effective recovery. To address these challenges, networking pods such as the API Gateway container will be employed.

Shared Services (Monitoring, Logging, Authentication, and Automation): Logging and monitoring are indispensable components of any containerized cluster, especially in a microservices architecture where numerous services are distributed across multiple machines or containers. The complexity of such deployments necessitates robust logging and monitoring solutions to swiftly diagnose issues. Additionally, each microservice requires its own security measures to manage access for team members, making authentication and authorization crucial considerations. Moreover, the implementation and maintenance of large systems like these would be challenging without automation tools. Automation mechanisms enable administrators to effectively implement and maintain the infrastructure, mitigating the risk of operational disasters.

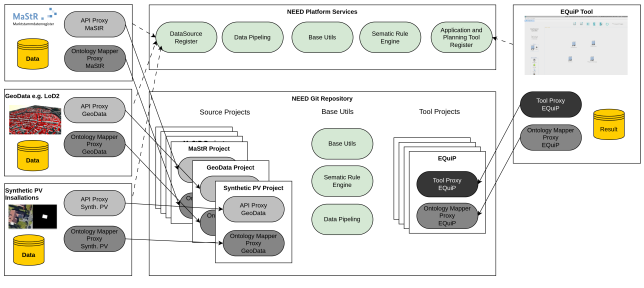

Fig. 8. Possible instantiation of the NEED ecosystem with API Proxy (light gray), Ontology Mapper (dark gray), Tool Proxy (black), Core Services and Base Utils (light green).

Figure 8 illustrates a possible instantiation of three data source providers (*Marktstammdatenregister*, LoD2 data, synthetic PV installations), which would register their capabilities at the Data Source Register (dashed arrow). This way, the Semantic Rule Engine may utilize the Ontology Mapper to integrate the new data sources. In parallel, the resources (API Proxy and Ontology Mapper) would be provided as Git projects. These may be deployed locally on the client side to enable the data access. A similar process would be conducted for tools. The requirements and data needs of the tools would be extracted via the Ontology Mapper and processed with the help of the knowledge graph. The result would provide the blueprint for the orchestration and preprocessing steps to retrieve the required data from the available data sources. The combined data records may then be forwarded to the Tool Proxy and finally processed by the tool. In this example, we would collect the already installed PV systems, update and crosscheck the result with synthetic PV installations, and calculate the still available roof area out of the LoD2 data to estimate the PV potential for a certain area.

7 SUMMARY AND OUTLOOK

In this work, we have emphasized the importance and need for energy-related and up-to-date data in order to perform reliable and repeatable planning tasks. In order to reduce the considerable effort required to collect and preprocess the necessary data, we will create an efficient, robust, and scalable infrastructure as part of the NEED research project. The project analyzed typical planning tasks and identified the requirements, stakeholders, and processes. The use cases developed for this were summarized in application scenarios and suitable case studies, which have already been implemented by the industry partners involved in the project, were documented for evaluation.

The idea of the project is to make numerous data sources available without replacing them, but rather integrating them. On the one hand, various conventional data sources were examined and their data and metadata evaluated with regard to the FAIR criteria. On the other hand, we want to close existing data gaps with the help of synthetic data generation approaches. In this context, we will also validate the quality of the developed methods, for example with available conventional data, in order to assess the transferability of the data. This will make it possible to transfer such data collection and provision to other countries that do not have corresponding data sources.

A major difficulty lies in combining different data for the different planning levels. Here the NEED project relies on ontologies or a knowledge graph. We believe that creating interoperability by using these technologies makes more sense and is more efficient than establishing a standard for data provision. Ultimately, the modules presented must be combined so that existing planning tools can access the required datasets. In some cases, further processing steps are required before the data can be merged and orchestrated. For this purpose, the project is developing a distributed service-based architecture, the essential components of which have been presented.

In the further course of the research project, the application scenarios presented will be implemented using conventional and synthetic data sources, which will be linked to the requirements of the planning tools using a knowledge graph, and the added value of the NEED ecosystem will be presented using the case studies. In this context, we will also evaluate the availability, scalability, and performance of the NEED platform. Overall, we are trying to make planning tasks more transparent, efficient, and uniform in order to further advance the energy transition.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support through the project executing agency Jülich (PTJ) with funds provided by the Federal Ministry for Economic Affairs and Climate Action (BMWK) due to an enactment of the German Bundestag under Grant No. 03EN3077A.

REFERENCES

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive science* 9, 1 (1985), 147–169.

Jethro Akroyd, Sebastian Mosbach, Amit Bhave, and Markus Kraft. 2021. Universal digital twin-a dynamic knowledge graph. *Data-Centric Engineering* 2 (2021), e14.

Christophe Bahim, Carlos Casorrán-Amilburu, Makx Dekkers, Edit Herczog, Nicolas Loozen, Konstantinos Repanas, Keith Russell, and Shelley Stall. 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal* 19 (Oct 2020). https://doi.org/10.5334/dsj-2020-041

Jiaru Bai, Kok Foong Lee, Markus Hofmeister, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2024a. A derived information framework for a dynamic knowledge graph and its application to smart cities. *Future Generation Computer Systems* 152 (2024), 112–126.

Jiaru Bai, Sebastian Mosbach, Connor J Taylor, Dogancan Karan, Kok Foong Lee, Simon D Rihm, Jethro Akroyd, Alexei A Lapkin, and Markus Kraft. 2024b. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nature Communications* 15, 1 (2024), 462.

Rukshan Batuwita and Vasile Palade. 2010. Efficient resampling methods for training support vector machines with imbalanced datasets. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

Rasoul Behravesh, Estefanía Coronado, and Roberto Riggio. 2019. Performance Evaluation on Virtualization Technologies for NFV Deployment in 5G Networks. In *2019 IEEE Conference on Network Softwarization (NetSoft)*. 24–29. https://doi.org/10.1109/NETSOFT.2019.8806664

Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. 2011. The MASTRO System for Ontology-Based Data Access. *Semantic Web* 2, 1 (2011), 43–53. https://doi.org/10.3233/SW-2011-0029

Arkadiusz Chadzynski, Nenad Krdzavac, Feroz Farazi, Mei Qi Lim, Shiying Li, Ayda Grisiute, Pieter Herthogs, Aurel von Richthofen, Stephen Cairns, and Markus Kraft. 2021. Semantic 3D City Database—An enabler for a dynamic geospatial knowledge graph. *Energy and AI* 6 (2021), 100106.

Arkadiusz Chadzynski, Shiying Li, Ayda Grisiute, Feroz Farazi, Casper Lindberg, Sebastian Mosbach, Pieter Herthogs, and Markus Kraft. 2022. Semantic 3D City Agents—An intelligent automation for dynamic geospatial knowledge graphs. *Energy and AI* 8 (2022), 100137. https://doi.org/10.1016/j.egyai.2022.100137

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

Fida Dankar and Mahmoud Ibrahim. 2021. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences* 11 (02 2021), 2158. https://doi.org/10.3390/app11052158

Valerie Eveloy and Wasiq Ahmed. 2022. Evaluation of Low-Carbon Multi-Energy Options for the Future UAE Energy System. *Sustainable Energy Technologies and Assessments* 53 (Oct. 2022), 102584. https://doi.org/10.1016/j.seta.2022.102584

Feroz Farazi, Nenad B Krdzavac, Jethro Akroyd, Sebastian Mosbach, Angiras Menon, Daniel Nurkowski, and Markus Kraft. 2020. Linking reaction mechanisms and quantum chemistry: An ontological approach. *Computers & Chemical Engineering* 137 (2020), 106813.

Alvaro Figueira and Bruno Vaz. 2022. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* 10, 15 (2022), 2733.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20 (1995), 197–243.

Hauke T.J. Henke, Francesco Gardumi, and Mark Howells. 2022. The Open Source Electricity Model Base for Europe - An Engagement Framework for Open and Transparent European Energy Modelling. *Energy* 239 (Jan. 2022), 121973. https://doi.org/10.1016/j.energy.2021.121973

Markus Hofmeister, Jiaru Bai, George Brownbridge, Sebastian Mosbach, Kok F Lee, Feroz Farazi, Michael Hillman, Mehal Agarwal, Srishti Ganguly, Jethro Akroyd, et al. 2024a. Semantic agent framework for automated flood assessment using dynamic knowledge graphs. *Data-Centric Engineering* 5 (2024), e14.

Markus Hofmeister, George Brownbridge, Michael Hillman, Sebastian Mosbach, Jethro Akroyd, Kok Foong Lee, and Markus Kraft. 2024b. Cross-domain flood risk assessment for smart cities using dynamic knowledge graphs. *Sustainable Cities and Society* 101 (2024), 105113.

Markus Hofmeister, Kok Foong Lee, Yi-Kai Tsai, Magnus Müller, Karthik Nagarajan, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2024c. Dynamic control of district heating networks with integrated emission modelling: A dynamic knowledge graph approach. *Energy and AI* 17 (2024), 100376.

Taeho Jo and Nathalie Japkowicz. 2004. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 40–49.

Rick Kazman, Phil Bianco, Sebastián Echeverría, and James Ivers. 2022. *Robustness*. Technical Report TR-004. Carnegie Mellon University. DOI: 10.1184/R1/16455660.

James Keirstead, Mark Jennings, and Aruna Sivakumar. 2012. A Review of Urban Energy System Models: Approaches, Challenges and Opportunities. *Renewable and Sustainable Energy Reviews* 16, 6 (Aug. 2012), 3847–3866. https://doi.org/10.1016/j.rser.2012.02.047

Aleksandar Kondinski, Angiras Menon, Daniel Nurkowski, Feroz Farazi, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2022. Automated rational design of metal–organic polyhedra. *Journal of the American Chemical Society* 144, 26 (2022), 11713–11728.

Ling Liu and M. Tamer Özsu (Eds.). 2018. *Encyclopedia of Database Systems*. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-8265-9

Clement Lork, Vishal Choudhary, Naveed Ul Hassan, Wayes Tushar, Chau Yuen, Benny Kai Kiat Ng, Xinyu Wang, and Xiang Liu. 2019. An Ontology-Based Framework for Building Energy Management with IoT. *Electronics* 8, 5 (April 2019), 485. https://doi.org/10.3390/electronics8050485

Zhongliang Lyu, Hua Wei, Xiaoqing Bai, and Chunjie Lian. 2020. Microservice-Based Architecture for an Energy Management System. *IEEE Systems Journal* PP (04 2020), 1–12. https://doi.org/10.1109/JSYST.2020.2981095

Michael Metzger, Mathias Duckheim, Marco Franken, Hans Joerg Heger, Matthias Huber, Markus Knittel, Till Kolster, Martin Kueppers, Carola Meier, Dieter Most, Simon Paulus, Lothar Wyrwoll, Albert Moser, and Stefan Niessen. 2021. Pathways toward a Decarbonized Future—Impact on Security of Supply and System Stability in a Sustainable German Energy System. *Energies* 14, 3 (2021). https://doi.org/10.3390/en14030560

Giota Paparistodimou, Alex Duffy, Robert Ian Whitfield, Philip Knight, and Malcolm Robb. 2020. A network science-based assessment methodology for robust modular system architectures during early conceptual design. *Journal of Engineering Design* 31, 4 (2020), 179–218.

Laura Pascazio, Simon Rihm, Ali Naseri, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2023. Chemical species ontology for data integration and knowledge discovery. *Journal of Chemical Information and Modeling* 63, 21 (2023), 6569–6586.

Matteo Giacomo Prina, Lorenzo Fanali, Giampaolo Manzolini, David Moser, and Wolfram Sparber. 2018. Incorporating Combined Cycle Gas Turbine Flexibility Constraints and Additional Costs into the EPLANopt Model: The Italian Case Study. *Energy* 160 (Oct. 2018), 33–43. https://doi.org/10.1016/j.energy.2018.07.007

Marco Pritoni, Drew Paine, Gabriel Fierro, Cory Mosiman, Michael Poplawski, Avijit Saha, Joel Bender, and Jessica Granderson. 2021. Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis. *Energies* 14, 7 (April 2021), 2024. https://doi.org/10.3390/en14072024

Hou Yee Quek, Markus Hofmeister, Simon D Rihm, Jingya Yan, Jiawei Lai, George Brownbridge, Michael Hillman, Sebastian Mosbach, Wilson Ang, Yi-Kai Tsai, et al. 2024. Dynamic knowledge graph applications for augmented built environments through "The World Avatar". *Journal of Building Engineering* 91 (2024), 109507.

Álvaro Sicilia and Gonçal Costa. 2017. Energy-Related Data Integration Using Semantic Data Models for Energy Efficient Retrofitting Projects. In *The Sustainable Places 2017 (SP2017) Conference*. MDPI, 1099. https://doi.org/10.3390/proceedings1071099

Heidi Silvennoinen, Arkadiusz Chadzynski, Feroz Farazi, Ayda Grišiūtė, Zhongming Shi, Aurel von Richthofen, Stephen Cairns, Markus Kraft, Martin Raubal, and Pieter Herthogs. 2023. A semantic web approach to land use regulations in urban planning: The OntoZoning ontology of zones, land uses and programmes for Singapore. *Journal of Urban Management* 12, 2 (2023), 151–167.

Mehmet Söylemez, Bedir Tekinerdogan, and Ayça Kolukısa Tarhan. 2024. Microservice reference architecture design: A multi-case study. *Software - Practice and Experience* 54, 1 (Jan. 2024), 58–84. https://doi.org/10.1002/spe.3241

Francois Spoto, Omar Sy, Paolo Laberinti, Philippe Martimort, Valerie Fernandez, Olivier Colin, Bianca Hoersch, and Aime Meygret. 2012. Overview Of Sentinel-2. In *2012 IEEE International Geoscience and Remote Sensing Symposium*. 1707–1710. https://doi.org/10.1109/IGARSS.2012.6351195

Dan Tran, Laura Pascazio, Jethro Akroyd, Sebastian Mosbach, and Markus Kraft. 2024. Leveraging text-to-text pretrained language models for question answering in chemistry. *ACS omega* 9, 12 (2024), 13883–13896.

Carolyn Whitnall, Elisabeth Oswald, and Luke Mather. 2011. An Exploration of the Kolmogorov-Smirnov Test as Competitor to Mutual Information Analysis. Cryptology ePrint Archive, Paper 2011/380. https://eprint.iacr.org/2011/380

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (15 Mar 2016), 160018. https://doi.org/10.1038/sdata.2016.18

Jim Young, Patrick Graham, and Richard Penny. 2009. Using Bayesian networks to create synthetic data. *Journal of Official Statistics* 25, 4 (2009), 549–567.

Xiaochi Zhou, Daniel Nurkowski, Angiras Menon, Jethro Akroyd, Sebastian Mosbach, and Markus Kraft. 2022. Question answering system for chemistry—A semantic agent extension. *Digital Chemical Engineering* 3 (2022), 100032.

Xiaochi Zhou, Daniel Nurkowski, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. 2021. Question answering system for chemistry. *Journal of Chemical Information and Modeling* 61, 8 (2021), 3868–3880.