# SciGRID_gas: The raw LKD data set

***Release 1.0***

**J.C. Diettrich & A. Pluta & W. Medjroubi**

**Jul 17, 2020**

# CONTENTS

*Impressum*

DLR Institute for Networked Energy Systems
Carl-von-Ossietzky-Str. 15
26129 Oldenburg
Germany
Tel.: +49 (441) 999 060

# LIST OF FIGURES

# LIST OF TABLES

**Summary**

The goal of SciGRID_gas is to develop methods to create an automated process that can generate a gas transmission network data set for Europe. Gas transmission networks are fundamental for simulations by the gas transmission modelling community, to derive major dynamic characteristics. Such simulations have a large scope of application, for example, they can be used to perform case scenarios, to model the gas consumption, to minimize leakages and to optimize overall gas distribution strategies. The focus of SciGRID_gas will be on the European transmission gas network, but the principal methods will also be applicable to other geographic regions.

Data required for such models are the gas facilities, such as compressor stations, LNG terminals, pipelines, etc. One needs to know their locations, in addition to a large range of attributes, such as pipeline diameter and capacity, compressor capacity, configuration, etc. Most of this data is not freely available. However throughout the SciGRID_gas project it was determined, that data can be found and grouped into two fundamental different groups: a) OSM data, and b) non-OSM data. The OSM data consists of geo-referenced facility data that is stored in the OpenStreetMap (OSM) data base, and is freely available. However, the OSM data set currently contains hardly any other information than the location of the facilities. The Non-OSM data set can fill some of those gaps, by supplying information such as pipeline diameter, compressor capacity and more. Part of the SciGRID_gas project is to mine and collate such data, and combine it with the OSM data set. In addition heuristic tools are required to fill data gaps, so that a complete gas network data set can be generated.

Here, this document describes one of the non-OSM data set, called the "LKD" data set, which originated from "Electricity, Heat, and Gas Sector Data for Modeling the German System" [LKD_MainRef]. This document here explains the origin and structure of this single data sets.

In this document, the chapter "Introduction" will supply some background information on the SciGRID_gas project, followed by the chapter "Data structure", that gives a detailed description of the data structure that is being used in the SciGRID_gas project. Chapter "Data sources" describes the LKD data set.

The appendix contains a glossary, references, location name alterations convention and finishes with the table of country abbreviation.

CHAPTER

# ONE

# INTRODUCTION

CHAPTER

TWO

# DATA STRUCTURE

A well designed and documented data structure is fundamental in any large scale project. Good data structure in combination with tools, based on algorithms, improve the performance of any project output.

This structure needs to represent the gas flow facilities as good as possible, Hence, it needs to include components, such as pipelines, compressors, etc. A finite number of components have been identified, that are required as building blocks of a gas network. In addition each component will contain attributes, such as pipeline diameter, maximal operating pressure, maximal capacity, number of turbines etc.

It is anticipated, that the adopted data structure can be implemented in different types of gas flow models and will be used by the research community for topics, such as sector coupling or identifying gas transmission bottlenecks.

Within the SciGRID_gas project, the structure of the data model is part of classes defined within the Python code. Alterations may occur over the duration of the project, but it is envisaged, that those will be small, and that compatibility will be assured.

The goal of this section is to describe in details the data structure that has been adopted and implemented into the Python code. This will be important in understanding other aspects of this document, such as exporting the data into CSV files or generating missing values.

Prior to the description of the data structure, the overall pathway of the data flow within the SciGRID_gas project will be explained, as it is believed, that such overview will help the reader.

## 2.1 Data structure description

This section contains information about the SciGRID_gas data structure, the format, and the code that can be used to import publicly available data into the project, so that it can be used in subsequent steps. Paramount for an understanding of the data structure is a good understanding of the terminology used throughout this section and the document in general. Hence, terminology will be introduced in the following sub-section.

### 2.1.1 Terminology

Throughout this document certain terms will be used, which will be described below and summarized as a picture in Figure 2.1.

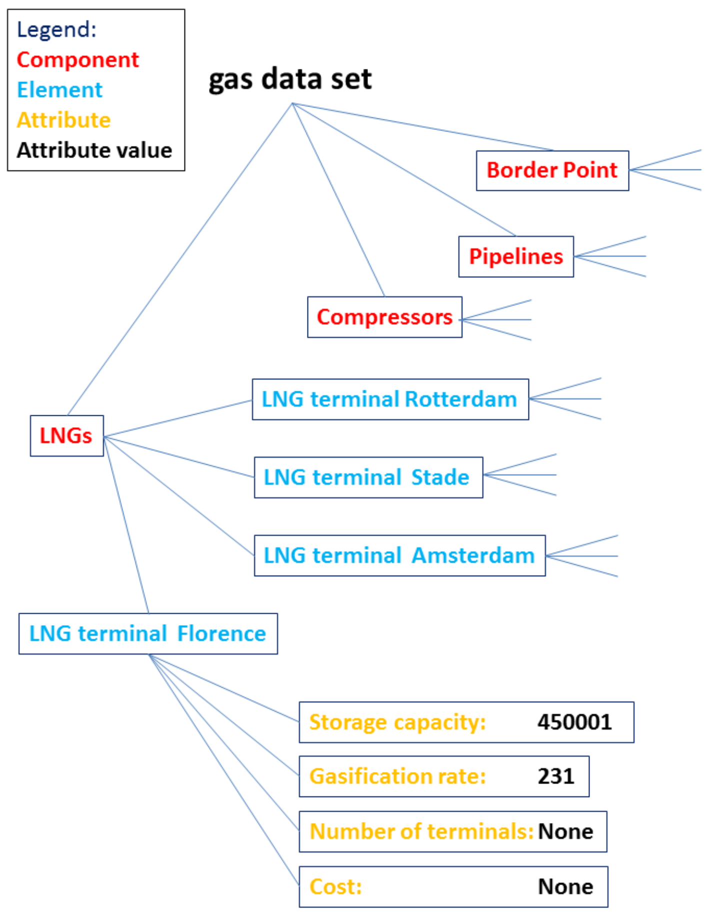


Figure 2.1: Data structure for the SciGRID_gas data set

### Gas transmission network

The term "gas transmission network" describes the physical gas transmission grid. This does not include the distribution of gas through gas distribution companies, but includes the long distance transmission of gas from producer countries to consumer countries, as carried out by the Transmission System Operators (TSO) [Wik20g].

### Gas component data set

The term "gas component data set" is used for all raw data of objects/facilities that have been loaded using SciGRID_gas tools into a Python environment. Gas component data sets are used as input into our SciGRID project. Several data sources can be loaded as gas component data sets, and then combined into a single gas component data set. However, not all elements (e.g. compressors) must be connected to pipelines, Hence, such a data set is referred to as a "gas component data set", and the emphasis is on the term **component**.

### Gas network data set

A "gas component data set" can be converted into a "gas network data set", by connecting all non-pipeline elements to nodes and all nodes are connected to pipelines, and as part of the process all network islands have been connected or removed, resulting in a single network. Therefore the network contains nodes and edges which are coherently connected, and all objects with the exception of pipelines are associated with nodes in this network, whereas pipelines are associated with edges. Hence, the emphasis here is on the term **network**.

### Component

There are several component types in a gas transmission network, such as compressors, LNG terminals, or pipelines. In Figure 2.1 they are coloured red. Hence, whenever the word "component" is mentioned, it refers to one of these components. There are roughly a dozen different components that will form a gas network data set. They will be briefly explained below.

### Element

The term "element" refers to individual facilities, e.g. the LNG Terminal in Rotterdam, or the compressor in Radeland. In Figure 2.1 they are coloured blue. The first one is an element of the component LNG terminals, whereas the second one is an element of the component compressors. Hence, many elements make up a component. However, all elements are referring to different facilities by default. This means in a single network, one cannot have two elements of a component describing the same facility. The structure of elements is described below.

### Attribute

"Attribute" is a term that is being used for the individual labels of the values that are associated with the elements. Examples for this term are gas "pipeline diameter", "maximum capacity", "max gas pipeline pressure", to name just a few and in Figure 2.1 they are coloured yellow. Overall there will be several hundred attributes in the SciGRID_gas project. However, the same attributes can occur in more than one component, e.g. "max flow capacity" exist for pipelines and also for compressors. Throughout the project, we have tried to keep the units of such attributes the same, so that there is no unit conversion required.

### Attribute value

Each attribute has a value, most likely a number or a string. In Figure 2.1 they are coloured black. While boolean (*True*/*False*) is also allowed, more likely a "1" will stand for *True* and "0" for *False*. However, some attribute values might not be given in the data source, therefore a no value attribute value does exist. In the Python code it is termed *None*.

The Figure 2.1 depicts the relationships between the terms "gas data set", "component", "element", "attribute", and "attribute value". As can be seen, a single gas data set consists of several components, where each component contains several elements, and each element has several attributes, which each come with a value, where "None" stands for unknown value. The heuristic processes described in this document at a later stage will fill those "None" values with generated values.

### Gas component types

A gas transmission network consists of different components, such as pipelines, compressors, etc. For the SciGRID_gas project a hand-full of components have been implemented, and will be described here briefly:

- *Nodes*: In a gas network, gas flows from one point to another point, which are given through their coordinates. All elements of all other components (such as compressor stations and power plants) have an associated node, which allows for the geo-referencing of each element. Overall the term "nodes" will be used throughout this document, as it aligns with graph theory aspects.
- *PipeLines*: *PipeLines* are one of the main components of the gas pipeline network. *PipeLines* allow for the transmission of the gas from one node to another. However, each pipe is unique. They might have different diameter, capacity or max pressure. In addition, a single *PipeLine* can connect several nodes. Therefore it could go from "Radeland" to "Bottrub" and then follow on to "Frankfurt". However, *PipeLines* do not need to connect more than 2 nodes, but can. The order of those nodes is important, and indicates the flow direction.
- *PipeSegments*: *PipeSegments* are almost identical to *PipeLines*, However, are only allowed to connect two nodes. Thus they have one start node and one end node, and are not passing via other nodes or other component elements in between, such as compressors or LNG terminals. Hence, any pipeline can easily be converted to multiple pipe-segments.
- *Compressors*: *Compressor* represent compressor stations, which are important. Gas travelling through the gas pipeline loses pressure due to friction on the pipeline walls and other factors. This will reduce the throughput of the gas amount. Hence, every so often (~ every 150 km), a compressor station is required, which increases the pressure of the gas, and Hence, allows the gas to flow through the gas pipeline. A gas compressor station contains several gas compressors units (turbines). Knowing the individual gas turbines is of an advantage, as those turbines can be combined in different ways, such as in series, or parallel, or combinations of those two options.
- *LNGs*: (LNG terminals and LNG storages) Some of the gas, which is being used throughout Europe, is supplied via ships to LNG terminals and LNG storage facilities. (From here onwards the acronym "*LNGs*" will be used instead "LNG terminals and LNG storage facilities".) As the transmission of gas would be extremely inefficient due to its volume, the gas state is changed to the liquid form (LNG gas), and then shipped. Ships arriving in Europe need special LNG terminals that can store LNG gas and subsequently re-gasify it. The storage and re-gasification of the gas are combined in the *LNGs* component and need to be part of any gas network for Europe.
- *Storages*: Part of the gas network will be gas storages. Gas storages are being used as gas pipeline capacities or gas production capacities might not be able to cover high demand periods, such as during the winter. Hence, large gas storage units are being filled during the summer periods while the overall demand is low, and if capacities of net supply allow it. This gas is then used during the winter period, and can compensate for shortcomings of the gas network or gas supply. Almost every country has their own gas storage units, ranging from smaller units to compensate for daily fluctuations to larger units, which compensate seasonal fluctuations. For the SciGRID_gas model the larger seasonal storage units are of more importance than the smaller ones,

as we are interested in the transmission gas pipeline network. However, any gas storage can be added and implemented into the gas network data set.

- *Consumers*: Part of the gas pipeline network is the knowledge of gas demand. Gas is added to the network at LNG terminals and European boundary cross border points. One type of users is the gas power plants. These can be added to the SciGRID_gas model, as this will specify local gas demand. In addition other consumers, such as city gas providers and large industries can also be added to the network data set.
- *Production*: These can be wells inside a country where gas is pumped out of the ground. Most of the gas used in Europe comes from outside of the EU, However, there are several smaller gas production sites scattered through Europe.
- *BorderPoints*: BorderPoints are cross border points (between different countries), which are mostly for the purpose of accounting the gas flow. Most large gas pipelines have cross border stations, e.g. Ellund (lat/long: 54.80181, 9.289079) at the border between Germany and Denmark, with gas facilities on both sides of the border.
- *EntryPoints*: These are special border points, as they are at the borders of the European Union and will be the gas entry points for the SciGRID_gas model data set.
- *InterConnectionPoints*: These are points between gas transmission operators, and will be found mainly within Europe, in particular at country borders. However, they can also be found within a single country, if there is more than one gas transmission operator.

### Element structure

As described above, elements are describing individual facilities, such as compressors or LNG terminals. However, the overall structure of those elements is the same for all elements of all components. The overall structure of those elements is described in the following part:

- *id*: A string, that is the ID of the element, and must be unique.
- *name*: A string that is the name of the element, such as "Compressor Radeland".
- *source_id*: A list of strings that are the sources of the element. As several elements from different sources could have been combined in a single element, one might need to know which are the original ids of the original sources.
- *node_id*: This is the ID of a geo-referenced point to which an element of the network is associated to. For a compressor, this will be just a single node_id, However, for a gas pipeline, that starts at one point and finishes at a different point, this entry would be a list of at least two node_id values.
- *lat*: This is the latitude value of an element. For pipelines, lat is a list of latitude values if known. The geo-referenced projection of the element that is being used in the SciGRID_gas project is: World Geodetic system 1984 (epsg:4326).
- *long*: The longitude, analogue to lat.
- *country_code*: This is a string indicating the official 2-digit country code (Alpha-2 code, see Chapter 5.4 for list of countries and their code). It represents the location of the element. As pipelines can pass through more than one country, the country code for pipes is the list of country codes of the countries the pipeline is passing through.
- *comment*: This is an arbitrary comment that is associated with the element.
- *tags*: This dictionary is reserved for OpenStreetmap data. It contains all associated key:value-pairs of an OpenStreetMap item.

In addition, there are three further groups of attributes to each element. Throughout the SciGRID_gas project, they have been coded as "dictionaries". They are called:

- *param*
- *method*
- *uncertainty*.

The structure within each dictionary is the same, However, their meaning is different. First of all the dictionary *param* (short for "parameter") contains a list of attributes and their values. This list of attributes will be different for each component. For the component *PipeLines* they might be pipeline diameter, max pipeline pressure, and max pipeline capacity. For the component *Compressors* they might be , such as number of turbines, overall turbine power, energy source of turbine and more.

So the other two attribute dictionaries are *method* and *uncertainty*. Each of those two dictionaries contains exactly the same list of attributes as the "param" dictionary. However, their attribute values reflect the name of the dictionary. E.g. the attributes in the dictionary *method* contain the information on the method used to derive the attribute value that is stored in the param dictionary. Here methods of value generation can include heuristic methods names (in form of strings) that have been implemented in the SciGRID_gas project. However, if attribute values are not being generated by the SciGRID_gas project, but originate from one of the input data sources, then the attribute values in the *method* dictionary is set to "raw".

Similar is the content of the *uncertainty* dictionary. It contains information on the uncertainty of the attributes from the *param* dictionary of that component. Again all attributes listed in the *param* dictionary are also present in the *uncertainty* dictionary. The attribute values here reflect the uncertainty of the attribute. Here, it is assumed, that attributes with a method of "raw" have an uncertainty of zero. Only for those attributes, which were generated during heuristic SciGRID_gas methods an uncertainty larger than zero will be specified.

## 2.2 Summary

The SciGRID_gas software is designed to construct a gas transmission network data set form different open source gas component data sets. The gas transmission data set needs to be available and stored in a precise and predefined way, which was described in this section. We have identified several *component*-types of a gas transmission network grid, like pipelines, compressor stations, LNG-terminals, etc. Each specific facility that falls under such a component is considered an *element* of that component. Each element is described by a list of *attributes* and correspondent *attribute values*.

CHAPTER

# THREE

# DATA SOURCES

Two thirds of the gas used in Europe is imported from non-EU states, and all gas required for the consumption needs to be distributed through the existing gas transmission pipelines in Europe. In the future gas consumption might rise, leading to additional pressure on the current infrastructure. In addition, gas facilities could play a vital role in reducing $CO_2$ emission, as excess electricity could be converted to gas, that could be stored and transmitted throughout Europe with the existing gas network. Hence, a reliable network data set for the European transmission network is essential. The data required for such models ranges from pipeline diameter, gas pressure within the pipeline, actual pipeline length, pipeline capacity, and underground storage volume to name just a few.

However, such data is the property of the transmission system operators (TSOs) and is therefore generally not freely available in the form and depth that is required for modelling purposes. The major reason for the difficulty of obtaining of such data is that most of the gas network infrastructure, namely pipelines, is buried underground. Thus a pipeline diameter is hard to estimate locally. In addition, almost all of the data is commercially sensitive.

However, there is a public drive to gather such data and subsequently make it available. The major platform through which this is occurring is the Open Street Map database [D18]. OSM is a geo-referenced database through which people can supply geo-referenced information on all man-made and natural structures, ranging from mountains to buildings. To achieve this, people throughout the world wander the globe and geo-reference everything that they can find. This also includes gas-pipeline markers, compressor stations or LNG terminals. However, the major problem remains that one cannot measure or estimate the diameter of the underground pipelines, or the number and size of the compressor turbines, as compressors are within buildings, which are fenced off. Hence, such information is hardly supplied to the OSM platform.

Nevertheless, some data is made available by gas transmission network operators, through different channels. E.g. information on the size and number of compressors could be made public through a press release, as part of a refurbishment. An example is given below (https://www.maz-online.de/Lokales/Teltow-Flaeming/Neue-Verdichterstation-entsteht-in-Radeland):

> "Die Eugal-Pipeline dient dazu, Gas aus der neuen Ostseepipeline Nord Stream 2 bis zur tschechischen Grenze zu leiten. 275 Kilometer von ihr verlaufen in Brandenburg. Grundsätzlich soll die neue Leitung parallel zur bestehenden Opal-Pipeline gebaut werden."

In addition some information can be found on company web pages, (https://www.open-grid-europe.com/cps/rde/SID-752BB6B5-E0A975F2/oge-internet-preview/hs.xsl/NewsDetail.htm?rdeLocaleAttr=en&newsId=50190C3B-E14F-4685-9E64-E40EEAB57A28):

> Open Grid Europe (OGE) is investing roughly EUR 150 million at its compressor station in Werne to improve the security and flexibility of energy supply for North Rhine-Westphalia and Germany. The upgrade of the station, which is one of the hubs of the pipeline network, will allow gas flows to be switched (reversed) from north to south and south to north. In addition, OGE is preparing the station for the upcoming transition from L- to H-gas. Through this fitness programme, the station's transmission capacity will increase by about 500,000 to 6.5 million $m^3$/h, which is equivalent to the annual consumption of more than 2,100 single-family homes. The project, which is due for completion at the end of 2018, is fully on track."

The data available can be separated into two different groups:

- OSM data: Data can be found in the OSM data base. OSM data is well geo-referenced, but contains little meta-information (information on the facility attributes, such as pipeline diameter or pipeline capacity). OSM data is very helpful to obtain accurate routes of pipelines.
- Non-OSM data: Non-OSM data have in general lower geographical accuracy but contain a lot of meta-information. Unfortunately, such information is only known for a few facilities. One exception to this rule are shapefiles from TSOs. They are rare, but well geo-referenced. However, the resolution of the meta information can vary from TSO to TSO.

One of the main challenges for SciGRID_gas is that, gas transmission data is incomplete and accumulated from different sources. Also such different sources can have different properties for one and the same facility. Hence, it is important to know, which data set supplies which information. Hence, this chapter here will introduce the relevant data sets (e.g. INET), starting off with the components, the elements for each component and then the attributes for each element.

## 3.1 Non-OSM data

Non-OSM data includes data from internet research, TSO press releases, TSO transparency platform, TSO public data, national open-source gas network data sets[1], etc.

Some of the TSO information had to be made available due to EU-regulations. Other information has been made public as part of a company's self presentation and advertisement. The information used by the SciGRID_gas project focuses on:

- the quality of the data
- the format of the data
- the level of representation of the data
- and the copyright restrictions on the data.

In addition, each data source is unique. Source specific tools need to be developed, so that all data sources can be made accessible for the SciGRID_gas project in the format as described in later chapter releases.

A significant portion of the project was spent on finding non-OSM data sets . Further data sources might be available, but unknown to the authors. If the authors are made aware of additional sources, the project will try to incorporate those, as this would only increase the depth of the data available and increase the applicability of the gas network data set and model.

Non-OSM data sources are very specific, addressing only certain aspects of the entire gas infrastructure. E.g. the GIE [GasIEurop20] data set supplies information on the daily gas flow in and out of gas storages in LNG terminals. However, they fall short on specifying the fundamental information of the actual physical location. Other data sets, such as the LKD [FMWP+17] data set is quite detailed in respect of pipelines, compressors and consumptions, however, only available for Germany.

Hence, the main task is to look closely at each data source, distil which data attribute values can be used, how it can be downloaded and incorporated into our SciGRID_gas model, and identify the copyright restrictions on the data source.

Due to copyright regulations, there are roughly two groups of data:

- Non copyright restrictive data (N-CRRD): here the copyright does not restrict the download, use and distribution of the data.
- Copyright restrictive data (CRRD): here the data can be downloaded and used internally, but not re-distributed to others.

[1] An entire gas network data set is only available from the UK, see https://www.nationalgridgas.com/land-and-assets/network-route-maps'.

The following is a list of the data sources that will be used throughout the project and an indication into which group of copyright restriction they fall:

- **OSM** (https://www.openstreetmap.org) (N-CRRD)
- **GB** (https://www.nationalgridgas.com/land-and-assets/network-route-maps) (CRRD)
- **NO** (https://www.npd.no/en/about-us/information-services/available-data/map-services/) (N-CRRD)
- **LKD** (https://tu-dresden.de/bu/wirtschaft/ee2/forschung/projekte/lkd-eu) (N-CRRD)
- **ENTSOG** (https://transparency.entsog.eu/) (CRRD)
- **EMap** (https://www.entsog.eu/sites/default/files/2020-01/ENTSOG_CAP_2019_A0_1189x841_FULL_401.pdf) (CRRD)
- **GIE** (https://www.gie.eu/) (N-CRRD)
- **GSE** (https://www.gie.eu/index.php/gie-publications/databases/storage-database) (N-CRRD)
- **IGU** (https://www.igu.org/) (CRRD)
- **GasLib** (http://gaslib.zib.de/) (N-CRRD)
- **INET** (see `Refs_InternetData`) (N-CRRD).

Each data set and source comes with a different copyright regulation. The copyright can be rather non-restrictive (e.g. INET) or can be restrictive (IGU). It is attempted to use only freely available data, so that such data can be re-distributed. In more restrictive data cases (IGU, GB), it is not allowed to download the data and distribute it to others. However, it is allowed to let other potential users know of the location of such data and supply them with tools, that allow them to carry out the same data download and subsequent incorporation of the data into a gas network data set.

**Note:**

In case that other users are aware of other data sources, that might be useful to this project, please get in touch and supply us with a brief description of the data and the location of such data, so that additional tools can be developed to incorporate the data in this project. Please use the following email address: developers.gas(at)scigrid.de

## 3.2 The Long-term Planning and Short-term Optimization (LKD) data set

The **Long-term planning and short-term optimization** data set (**LKD**) is the second of two non-OSM data sets, that contain geo-referenced gas facilities. It was generated by several German research institutes and funded through the German government grants. It was part of a much larger research project (see link below). Here the gas facilities from the lk-DEU data set were used and incorporated into SciGRID_gas data model as the **LKD** data set. It contains information on gas pipelines, gas production sites, gas storages, compressor locations, and nodes.

As this data set is extremely well geo-referenced, it is of particular interest for the SciGRID_gas data project. The LKD data set can be used in conjunction with the OSM data set for training purposes, and as a data source for the heuristic processes, due to its depth of data attributes.

The LKD facilities data set came in form of a shapefiles, and consisted of polylines with some attributes, such as pipe diameter, max gas flow capacity and more. In addition parts of the shapefiles were tables of facilities, with information on storages, production, and industrial demand. Great care was taken from the original data set producers, to create a data set with a vast amount of attributes, which will be used throughout the SciGRID_gas project. Overall the topological quality of the data set is good, as could be verified by doing sample checks. Gas sites could be found on satellite images within a few hundred meters. Due to its data depth, the entire LKD data set is being loaded by tools, and has been incorporated into the SciGRID_gas gas data project.

**Further external information on the lk-DEU data set**

More information on the data can be found under the following URL:

https://www.ewl.wiwi.uni-due.de/nl/forschung/forschungsprojekte-ewl/lkd-eu-langfristige-planung-und-kurzfristige-optimierung-des-elektrizitaetssystems-in-deutschland-im-europaeischen-kontext/.

This chapter describes the '(Long-term planning and short-term optimization of the German electricity system within the European' data set [FMWP+17]. The project was a joined effort by:

- German Institute for Economic Research (DIW Berlin)
- Working group for Infrastructure Policy (WIP) at Technische Universitaet Berlin (TUB)
- Chair of Energy Economics (EE2) at Technische Universitaet Dresden (TUD)
- House of Energy Markets & Finance at University of Duisburg-Essen.

This project was funded by the German Federal Ministry for Economic Affairs and Energy through the grant "LKD-EU", FKZ 03ET4028A, with the aim of presenting a status quo of the German energy sector. The following three energy media were part of their project:

- electricity
- heat
- natural gas.

### 3.2.1 Pre requirements for accessing the LKD data set

The SciGRID_gas project has received the right to use, change and redistribute the LKD data under an open license. However, if you use this data or any data set which incorporates this data, you are also obliged to cite the original authors of the LKD data as follows:

Kunz et al. 2017, Data Documentation: Electricity, Heat, and Gas Sector Data for Modeling the German System

In addition, the data set can be downloaded from the following location:

https://zenodo.org/record/1044463#.Xah7i2ZCSUk

Please put a copy into the following location:

/SciGRID_gas/Eingabe/LKD/

Further data re fixes of the LKD data set are required, However, will be supplied with the code, and have been implemented in the CSV LKD output data set.

### 3.2.2 Manipulation of the LKD data

The LKD gas facilities data set came in form of several shapefiles. It contained several tables, which were read in with tool and dissected to fit into the data structure of SciGRID_gas. The components that were read in are:

- *PipeSegments*, from the 'pipelines_utf8.shp' shapefile
- *Nodes*, from the 'nodes_utf8.shp' shapefile
- *Productions*, from the 'productions_utf8.shp' shapefile
- *Storages*, from the 'storages_utf8.shp' shapefile.

Subsequent to reading the data from the shapefiles, it was necessary to convert the data so that it adheres to the SciGRID_gas data structure. Some inconsistencies were found with the data set. The following fixes of the LKD data set had to be carried out:

- Some node ids were found more than once in the original data set for different nodes. Hence, this was manually rectified by changing some attributes of 29 nodes. This information was generate and subsequently supplied to the LKD tool via a CSV file.
- Some nodes had a wrong country code setting. Here again those nodes were determined manually, the correct country code was written to a CSV document, which gets read by the LKD tools as part of loading the data. For 10 nodes, the country code attribute needed to be changed.

In an additional step, the component Compressor was generated by using information that was supplied with the Nodes. The Node elements contained an attribute "comp_units", which stands for "number of compressor units". Hence, if this value was larger than 0, then it was assumed, that the node contained a single compressor element at that location. In addition, the attribute "comp_units" was then used as the value for the number of compressor turbines at the compressor location. E.g. if the value was two, then the compressor element's attribute "num_turb" was set to two.

At this stage of the LKD data manipulation process, there were more than 1800 pipe-segments with more than 1400 nodes. It was not apparent why there were so many nodes. For many pipe-segments, two individual pipe-segments that connect with the same node contained the same attributes with the same values, and the node in question only connected two pipe-segments, Hence, not forming a T-section. Hence, pipe-segments were joined and nodes removed if the following rules applied:

- the node in question connects only two pipe-segments
- the attributes values for 'max_pressure_bar', 'is_H_gas', 'diameter_mm', 'pipe_class_LKD' need to be identical for both pipe-segments. An exception is made for the node "Haidach" and "N_805129", where not pipe-segment joining took place.

This process reduced the number of segments to 1261, and the overall number of nodes to 938.

For some of the attribute values, the unit of the attribute value did not "agree" with the units used within the SciGRID_gas data project. Hence, unit transformation had to be carried out for the following attributes of the following components:

- component "Storages", attribute converted from 'max_cap_pipe2store_GWh_per_d' to 'max_cap_pipe2store_M_m3_per_d'
- component "Storages", attribute converted from 'max_cap_store2pipe_GWh_per_d' to 'max_cap_store2pipe_M_m3_per_d'

- component "PipeSegments", attribute converted from 'max_cap_GWh_per_d' to 'max_cap_M_m3_per_d'.

Subsequently, the old attributes with the wrong units were removed from the component elements.

Further attributes were added to the component "PipeSegments":

- the length of the pipe-segment was derived using the polylines of each pipe-segment
- the average latitude and average longitude was calculated by using the polylines of each pipe-segment.

In addition, the attribute "exact" was added to each Nodes element and a value of one was given.

In addition, for each Nodes element the following attributes were removed:

- 'compressor'
- 'ugs'
- 'production'
- 'comp_units'.

### 3.2.3 LKD data density

The data of the LKD data set contains elements from the following components:

- *PipeSegments*
- *Compressors*
- *Productions*
- *Nodes*
- *Storages*.

Each of those components and their attributes will be described below.

As all components have the following attributes, they are presented here ones:

- *id*: unique identifier
- *name*: name of the pipe-segment
- *source_id*: a source id
- *node_id*: the id of the start and the end node of the pipe-segment
- *lat*: a list of latitude values
- *longitude*: a list of longitude values
- *country_code*: a string pair indicating the country code of the start and the end point
- *comment*: a user comment.

### Pipe-segments

Overall there are 1261 *PipeSegments* elements in the LKD data set. In addition to the default attributes, the following non-standard attributes (see Table 3.1) are supplied. The number of attribute values supplied for each attribute is given by the parameter 'data density' (see Chapter 5.1):

Table 3.1: LKD *PipeSegments* data summary

| Attribute name | Description | Units | data density [%] |
|---|---|---|---|
| diameter_mm | a pipe diameter | mm | 100 |
| is_H_gas | the gas type identifier | 1 or 0 | 100 |
| length_km | the total distance of the pipe-segment | km | 100 |
| max_cap_M_m3_per_d | maximum gas flow capacity | $\mathrm{Mm^3d^{-1}}$ | 17 |
| max_pressure_bar | maximum allowed pressure in the gas pipe | bar | 100 |
| operator_name | operator name | | 99 |
| pipe_class_LKD | gas pipe-segment class type | 1 to 6 | 88 |
| lat_mean | calculated mean latitude value | degree | 100 |
| long_mean | calculated mean longitude value | degree | 100 |

#### pipe_class_LKD

For reasons of attribute generation at a later stage, the values for *pipe_class_LKD* have been converted from A, B, C,... to 1, 2, 3, ...

### Compressors

Overall there are 13 *Compressors* elements in the LKD data set. In addition to the default attributes, the following non-standard attributes are supplied (see Table 3.2) and partially populated for the component *Compressors*:

Table 3.2: LKD *Compressors* data summary

| Attribute name | Description | Units | data density [%] |
|---|---|---|---|
| entsog_key | key associated with EntsoG facility | | 38 |
| license | indicator of the license | | 100 |
| num_turb | the number of compressor turbines | | 100 |
| operator_name | name of the operator | | 100 |

### Storages

Overall there are 14 *Storages* elements in the LKD data set. In addition to the default attributes, the following non-standard attributes (see Table 3.3) are supplied and partially populated for the component *Storages*:

Table 3.3: LKD *Storages* data summary

| Attribute name | Description | Units | data density [%] |
|---|---|---|---|
| entsog_key | key associated with EntsoG facility | | 100 |
| max_cap_pipe2store_M_m3_per_d | maximum gas flow from the network into the storage unit | $\mathrm{Mm^3d^{-1}}$ | 100 |
| max_cap_store2pipe_M_m3_per_d | maximum gas flow from the storage unit into the network | $\mathrm{Mm^3d^{-1}}$ | 100 |
| operator_name | name of the operator | | 100 |

### Productions

Overall there are 6 *Productions* elements in the LKD data set. In addition to the default attributes, the following non-standard attributes (see Table 3.4) are supplied and partially populated for the component *Production*:

Table 3.4: LKD *Productions* data summary

| Attribute name | Description | Units | data density [%] |
|---|---|---|---|
| entsog_key | key associated with EntsoG facility | | 100 |
| max_production_M_m3_per_d | maximum gas production | $\mathrm{Mm^3d^{-1}}$ | 100 |
| is_H_gas | boolean re H gas type | 1 or 0 | 100 |
| operator_name | name of the operator | | 100 |

### Nodes

Overall there are 938 *Nodes* elements in the LKD data set. In addition to the default attributes, the following non-standard attributes (see Table 3.5) are supplied and partially populated for the component *Nodes*:

Table 3.5: LKD *Nodes* data summary

| Attribute name | Description | Units | data density [%] |
|---|---|---|---|
| crossborder | boolean if node is a gas cross border point | | 100 |
| entry | boolean if node is a gas entry point | | 100 |
| entsog_key | key associated with EntsoG facility | | 9 |
| exact | value indicating the accuracy in geo-referencing | 1 to 5 | 100 |
| exit | boolean if node is a gas exit point | | 100 |
| H_L_conver | boolean indicating of converter between H & L gas | | 100 |
| license | license key | | 100 |
| operator_Z | additional operator name | | 64 |
| operator_name | name of the operator | | 98 |

### Additional data from the LKD data set

In addition, there is data on gas demand on a spatial level of Nut-3, leading to 402 elements. However, currently this is not being used, However, might be used at a later stage.

## 3.2.4 Copyright and disclaimer for the LKD data set

Please see sub-section above for the copyright regulation for this LKD data set.

### 3.2.5 Summary LKD data

The gas pipeline and gas facility from the LKD data set is of great importance to the SciGRID_gas project. It is one of only two non-OSM data sets that contain gas facilities that are geo-referenced, and Hence, can be used for validation processes covering all of Germany. In addition, it contains some attribute values in respect of gas pipelines that are fundamental for the gas data model. This data set was available through a German research project and was downloadable from a web page. Tools have been written to load the LKD shapefiles and make them accessible throughout the SciGRID_gas project.

Below the Table 3.6 summarises the number of elements for each component found:

Table 3.6: LKD component summary

| Component Name | Count |
|---|---|
| BorderPionts | 0 |
| Compressors | 13 |
| ConnectionPoints | 0 |
| Consumers | 0 |
| EntryPoints | 0 |
| InterConnectionPoints | 0 |
| LNGs | 0 |
| Nodes | 938 |
| PipeSegments | 1261 |
| Production | 6 |
| Storages | 14 |

In addition, the map in Figure 3.1 visualizes the data for Germany.

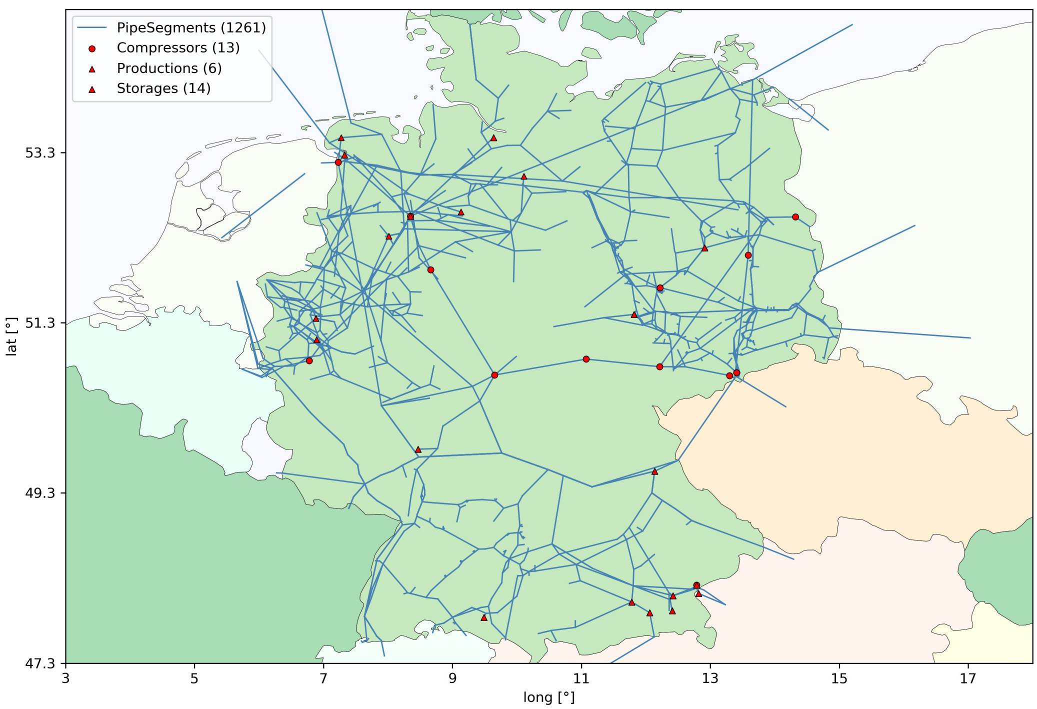


Figure 3.1: Map of components of the LKD data set.

## 3.3 Data summary

SciGRID_gas is based on open source data. To generate a gas pipeline network data set, one needs to access different data sets that were found throughout the project and presented here. Emphasis was given to depict the number of elements per component and the data density for each data set.

## 3.4 Summary

Gas component data sets come in different forms, licenses, formats and detail. The SciGRID_gas project can process such data and combine them to a consistent and reliable network data set.

The underlying gas component data sets were categorized into two different groups:

- OSM data: This is data originating from the OSM data base, containing well geo-referenced locations of gas facilities, such as pipe locations or gas storage facilities. However, it comes with very few meta information.
- Non-OSM data: These are all other data sources, which can "supply" detailed information on some of the gas facilities attributes. However, this information is sparse, as published only for a few facilities. Here, the INET data set was introduced as an example of the non-OSM data set, and the pathway of converting the raw data from the www into SciGRID_gas project component structure.

Here detailed information on one or several data sources have been given, and should be used as a reference for later data processes.

CHAPTER

# FOUR

# CONCLUSION

This document here is the documentation of one of the data sets that is part of the SciGRID_gas project. This document here started off with the introduction of the SciGRID_gas project, such as funding, duration and goals. In a subsequent chapter the data structure within the SciGRID_gas project was described, such as components, elements, attributes and attribute values, so that the transmission network data set could be an input to certain gas flow model. The third chapter introduced the LKD data set, which is a data set that generated by a third party, and the SciGRID_gas project is fortunat enough, that it can use the LKD data set and redistribut it through this project. Here only the raw LKD data set is being published, hence it will have some missing values. Despite some missing values, the re-distributed LKD data set consists of 0 LNG terminals, 14 storage facilities, 13 compressors and 1261 pipe segments throughout Germany.

# CHAPTER FIVE

# APPENDIX

## 5.1 Glossary

The glossary terms can be found in Table 5.1.

Table 5.1: Glossary

| Name | Abbreviation | Description | Units |
|---|---|---|---|
| InternetDaten data set | INET | This is the label/name for the InternetDaten data set | |
| Gas Infrastructure Europe data set | GIE | This is the label/name for the Gas Infrastructure Europe data set | |
| Gas Storage Europe data set | GSE | This is the label/name for the Gas Storage Europe data set | |
| Long-term planning and short-term optimization data set | LKD | This is the label/name for the Long-term planning and short-term optimization data set | |
| International Gas Union data set | IGU | This is the label/name for the International Gas Union data set | |
| Norwegian data set | NO | This is the label/name for the Norwegian data set | |
| EntsoG-Map data set | EMAP | This is the label/name for the EntsoG-Map data set | |
| component | | A gas network consists of different components, such as pipelines, compressors LNG terminals and more. However, for a gas transmission network, there is a handful of components only: pipeline, compressor, LNG terminal, storage, entry point, border point, connection point, consumer, node, and production | |
| element | | Elements are instances of component. Hence, we speak of 10 compressor elements, if we have a data set that has 10 compressors. Here then we can refer to the first or the last or any element of such component | |
| attribute | | Gas facilities, such as pipelines or compressors, can be described with a large number of parameters, such as pipeline diameter, or compressor capacity. Those parameters are referred to as attributes. Hence, each component has a list of properties, which are different from one component to another component | |
| facility | | General term used for a gas appliance, such as compressor element, or LNG terminal | |

Table 5.2: Glossary

| Name | Abbreviation | Description | Units |
|---|---|---|---|
| PipeLine | | This is a gas pipeline entity, which has one start and one end point, however can run via many nodes, compressors and other gas network elements | |
| PipeSegment | | This is a gas pipeline, that has only one start and one end point, but no nodes in-between, Hence, only goes from one node to another node | |
| LNG | LNG | Liquefied natural gas | |
| CNG | CNG | Compressed natural gas | |
| flow duration curve | FDC | It is the cumulative frequency curve that shows the percent of time specified flow were equal or exceeded during a given period. The information on occurrence of events is lost | |
| Energiewende | | German term for the change in using primary energies, the move away from coal to renewable energies, such as wind or solar | |
| gas component data set | | Raw input data, associated with components of the gas transmission grid | |
| gas network data set | | Output data, a coherent network of gas transmission components | |
| OSM | OSM | Data that is available from the openstreetmap.org | |
| non-OSM | Non-OSM | Data that is not part of the OSM data set | |
| gas type | | There are two types of gas High (H) and Low (L) calorific gas | |
| mean absolute error | MAE | mean difference between input values and estimated values | |
| data density | | This is the ratio of the number of usable (not missing) attribute values over number elements of the component | [%] |
| Transmission System Operators | TSO | This is an entity entrusted with the transportation of natural gas/electricity, as defined by the European Union | |
| gas transmission network | | This describes the physical gas transmission grid, however excludes any facilities/components that would be part of a distribution network and their facilities. This projects goal is to create an open source gas network data set that can be used to describe the European gas transmission network | |
| gas component data set | | The term "gas component data set" is used for all raw data of objects/facilities that have been loaded using SciGRID_gas tools into a Python environment. However, not all elements (e.g. compressors) must be connected to pipelines, Hence, such a data set is referred to as a "gas component data set", and the emphasis is on the term **component** | |
| gas network data set | | A "gas component data set" can be converted into a "gas network data set", by connecting all non-pipeline elements to nodes and all nodes are connected to pipelines, and as part of the process all network islands have been removed, resulting in a single network. Therefore the network contains nodes and edges which are connected, and all objects with the exception of pipelines are associated with nodes in this network, whereas pipelines are associated with edges. Hence, the emphasis here is on the term **network** | |

## 5.2 Unit conversions

Table 5.3: Unit conversions

| From Unit | To Unit | MultiVal |
|---|---|---|
| LNG Mt | LNG $\mathrm{Mm}^3$ | 2.47 |
| gas $\mathrm{tm}^3/\mathrm{h}$ | gas $\mathrm{Mm}^3/\mathrm{d}$ | 24/1000 |
| LNG $\mathrm{Mm}^3$ | gas $\mathrm{Mm}^3$ | 584 |
| LNG Mt | gas $\mathrm{Mm}^3$ | 1442.48 |

## 5.3 Location name alterations

Location names should be changed into the 26 letters used in the English language.

For names from the individual countries please follow the suggested approach:

- Germany/Austria: *Umlaute* to be replaced with the letter followed by an 'e', e.g.: ü = ue.
- France/Belgium: Omit accent de gues and accent de graphs, e.g.: ó = o.
- Sweden: Please change the last three letters of the Swedish alphabet and replace e.g.: ä = a.
- Poland: Please change any letter, that cannot be found in the English alphabet, knowing that for some letters, that one can only use a single letter instead of the three different letters used in the Polish alphabet, e.g.: z = z.
- Spain/Portugal: Please change any letter, that cannot be found in the English alphabet, e.g.: ñ = n.
- Greece: Please do not use Greek letters. Please try to write the Greek words with Latin letters.
- Denmark: Please change any letter that contains non-English letters, e.g.: "å" with "aa".
- Slovakia, Czech Republic, Hungary, Rumania, Latvia, Lithuania, Estonia, Bulgaria, Slovenia, Croatia: PLEASE use your common sense, based on the examples from the other countries above.

## 5.4 Country name abbreviations

For convenience we provide a short list of names and two-digit codes (see Table 5.4) for the probably most important countries associated with the European Transmission Grid.

Table 5.4: Country codes

| Country name | Country code | Country name | Country code |
|---|---|---|---|
| Albania | AL | Kosovo | XK |
| Armenia | AM | Latvia | LV |
| Austria | AT | Liechtenstein | LI |
| Azerbaijan | AZ | Lithuania | LT |
| Belarus | BY | Luxembourg | LU |
| Belgium | BE | Malta | MT |
| Bosnia and Herzegovina | BA | Moldova | MD |
| Bulgaria | BG | Montenegro | ME |
| Croatia | HR | Netherlands | NL |
| Cyprus | CY | Norway | NO |
| Czech | CZ | Poland | PL |
| Denmark | DK | Portugal | PT |
| Estonia | EE | Romania | RO |
| Finland | FI | Serbia | RS |
| France | FR | Slovakia | SK |
| Georgia | GE | Slovenia | SI |
| Germany | DE | Spain | ES |
| Greece | EL | Sweden | SE |
| Hungary | HU | Switzerland | CH |
| Iceland | IS | Turkey | TR |
| Ireland and Northern Ireland | IE | Belarus | UA |
| Italy | IT | Great Britain | GB |
| Russia Federation | RU | Europe | EU |

## 5.5 Acknowledgement

We acknowledge the contribution of Dr. Ontje Luensdorf from the DLR Institute of Networked Energy System to the SciGRID_gas project.

# BIBLIOGRAPHY


[B19] Sandvik B. World Borders. http://thematicmapping.org/downloads/world_borders.php, 2019. Accessed: 2019-07-07.

[D18] Helle D. OpenStreetMap - Deutschland. https://www.openstreetmap.de/, 2018. Accessed: 2019-12-12.

[FMWP+17] Kunz F., Kendziorski M., Schill W.-P., Weibezahn J., Zepter J., von Hirschhausen C., and Hauser P. *Electricity, Heat, and Gas Sector Data for Modeling the German System*. Deutsches Institut für Wirtschaftsforschung, Daten Documentation 92, Berlin, 2017.

[HL99] Alt H. and Guibas L. *Discrete geometric shapes: matching, interpolation, and approximation-a survey*. Sack JR, Urrutia J, Handbook of Computational Geometry, Elsevier, New York, NY, 1999.

[JCJAW20] Diettrich J.C., Dasenbrock J., Pluta A., and Medjroub W. SciGRID_gas: The filled INET data set. 2020. Accessed: 2020-07-08.

[MBTC14] Ahmed M, Fasy B.T., and Wenk C. *New Techniques in Road Network Comparison*. Penguin Random House, New York, NY, 2014.

[MDR+17] Schmidt M., Aßmann D., Burlacu R., Humpola J., Joormann I., Kanelakis N., Koch T., Oucherif D., Pfetsch M.E., Schewe L., Schwarz R., and Sirvent M. *GasLib—A Library of Gas Network Instances*. 2017. doi:{10.3390/data2040040}.

[RLP85] Graham R.L. and Hell P. On the history of the minimum spanning tree problem. *Annauls of the History of Computer*, 7(1):43–57, 1985. doi:{10.1145/2729977}.

[sl19] scikit-learn. 1.1. Linear Models (scikit learn). https://scikit-learn.org/stable/modules/linear_model.html, 2019. Accessed: 2019-08-08.

[UoO14] USA University of Oregon. Comparing distributions: Z Test. http://homework.uoregon.edu/pub/class/es202/ztest.html, 2014. Accessed: 2020-07-07.

[W16] Medjroubi W. SciGRID_power. https://power.scigrid.de, 2016. Accessed: 2019-09-09.

[Wik20a] Wikipedia. Bootstrapping (statistics). https://en.wikipedia.org/wiki/Bootstrapping_(statistics), 2020. Accessed: 2019-06-06.

[Wik20b] Wikipedia. Cross-validation (statistics). https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Exhaustive_cross-validation, 2020. Accessed: 2019-07-07.

[Wik20c] Wikipedia. Jackknife resampling. https://en.wikipedia.org/wiki/Jackknife_resampling, 2020. Accessed: 2019-08-08.

[Wik20d] Wikipedia. Lasso (statistics). https://en.wikipedia.org/wiki/Lasso_(statistics), 2020. Accessed: 2020-04-04.

[Wik20e] Wikipedia. Limited-memory BFGS. https://en.wikipedia.org/wiki/Limited-memory_BFGS, 2020. Accessed: 2020-06-06.

[Wik20f] Wikipedia. Out-of-bag error. https://en.wikipedia.org/wiki/Out-of-bag_error, 2020. Accessed: 2019-07-07.

[Wik20g] Wikipedia. Transmission system operator. https://en.wikipedia.org/wiki/Transmission_system_operator/, 2020. Accessed: 2019-09-09.

[Wik20h] Wikipedia. JAGAL. https://en.wikipedia.org/wiki/JAGAL, 2020. Accessed: 2020-01-01.

[Y13] Khalid Y. What is Pickle in python? https://pythontips.com/2013/08/02/what-is-pickle-in-python/, 2013. Accessed: 2019-10-10.

[BMWi11] BMWi. Forschung für eine umweltschonende, zuverlässige und bezahlbare Energieversorgung. https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/6-energieforschungsprogramm-der-bundesregierung.pdf?__blob=publicationFile&v=12, 2011. Accessed: 2019-02-02.

[BMWi20] BMWi. Home page of BMWi. https://www.bmwi.de/Navigation/DE/Home/home.html, 2020. Accessed: 2020-03-03.

[BundesregierungDeutschland20] Bundesregierung Deutschland. Home page of Bundesregierung Deutschland. https://www.bundesregierung.de/Webs/Breg/DE/Themen/Energiewende/_node.html, 2020. Accessed: 2020-01-01.

[GasIEurop20] Gas Infrastructure Europ. Home page of Gas Infrastructure Europ. https://agsi.gie.eu, 2020. Accessed: 2020-01-01.