Automatic Creation of Knowledge Graphs from IEC 61850-Tagged Datasets for Renewable Energy Systems

Lina Nachabe[1], Maxime Lefrançois[1], Cyril Effantin[2] and Antoine Zimmermann[1]

[1]*Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, Saint-Étienne, France*
[2]*EDF R&D, Palaiseau, France*

Abstract

The rapid growth of renewable energy toward industry 4.0, particularly renewable solar, has amplified the need for findable, accessible, interoperable, and reusable data sets collected from meteorological stations, sensors, renewable equipments, infrastructures, and others. In particular, solar plants generate large volumes of operational data, often structured using the IEC 61850 standard, a widely adopted protocol for energy systems that defines models for communication between intelligent devices. While IEC 61850 ensures consistency at the data exchange level, it does not provide full semantic interoperability needed for advanced analytics, data discovery, cross-domain integration, and automated reasoning. This paper describes an automatic pipeline for transforming IEC 61850-tagged data into knowledge graphs, enabling seamless integration with diverse datasets such as weather and grid structure for operation maintenance services as well as energy prediction services. This pipeline allows the transformation of these tags into a KG aligned with the Omega-X ontology, reducing the need for manual intervention, enabling continuous data integration, and enhancing the reuse of these data by service providers in the energy sector. We evaluate this approach in the context of the Omega-X project, using real-world datasets from a solar park that combines meteorological and electrical parameters.

Keywords

IEC 61850, knowledge graph construction pipeline, renewables solar ontology, solar datasets, FAIR,

1. Introduction

The continuous use of renewable energy systems has driven the emergence of new digital services designed to accelerate the integration of photovoltaic (PV) technologies. These services include the optimization of Operation & Maintenance (O&M) workflows for early fault detection, as well as congestion forecasting, and energy production estimation [1]. Such services rely on heterogeneous datasets collected from solar park equipment, weather stations, infrastructure systems, and other related sources. However, they are often developed using siloed datasets originating from individual companies or even specific departments within an organisation. As a result, the models and services built on these limited datasets may not be reused by other stakeholders and have limited performance when applied across diverse environments or varying operational contexts. Consequently, European Data Space (DS) targets to overcome these barriers by promoting data sharing across multiple stakeholders. In particular, the European Energy DS aims to exchange data between data providers and services providers to develop more resilient and adaptable services that enhance the performance of PVs [2]. This approach enables interoperability, innovation, and efficient energy management while maintaining data sovereignty. In the context of the European Energy DS, the Omega-X project [1] addressed semantic interoperability challenges across four use case families, including renewables, flexibility, local energy communities, and electro-mobility using ontologies and Knoweldge Graphs (KGs). In the context of the projects, datasets exchanged between data providers and service providers are semantified using

Posters & Demos Track, co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

lina.nachabe@emse.fr (L. Nachabe); maxime.lefrancois@emse.fr (M. Lefrançois); cyril.effantin@edf.fr (C. Effantin); antoine.zimmermann@emse.fr (A. Zimmermann)
00000-0003-3924-3336 (L. Nachabe); 0000-0001-9814-8991 (M. Lefrançois); 0000-0003-1502-6986 (A. Zimmermann)

[1] Omega-X - Orchestrating an interoperable sovereign federated Multi-vector Energy data space built on open standards and ready for GAia-X, funded by the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101069287 – https://omega-x.eu

the Omega-X Ontology [2], which models the structure of diverse energy datasets and services. In the renewable use case family, different data providers wanted to share solar energy datasets about pilot sites. Électricité de France (EDF) was one of them and wanted to share an heterogeneous set of CSV files that rely on the International Electrotechnical Commission IEC 61850 standard [3] and are generated on a weekly basis for a pilot site[3]. The IEC 61850 standard is part of the IEC Technical Committee 57 reference architecture for electric power systems, and defines a communication protocol for power utility automation. It defines a general tagging mechanism for data elements, specialised for several domains such as substation automation[4], hydropower plant, wind plant, solar plant, etc. [3]. Although standardised, IEC 61850 tags are difficult to interpret by non-domain experts, making it difficult to integrate datasets based on these tags with other datasets. In addition, the solar power datasets shared by EDF presented different structures with no explicit semantics, hindering both automated integration and cross-dataset analysis. This paper reports on a pipeline for creating KGs conformant to the Omega-X ontology from these IEC 61850-tagged datasets.

2. Automatic pipeline for KG creation

In IEC 61850 tags, each physical device (Ph, in this case the solar park) consists of logical devices (LD), which consist of logical nodes (LN) that represent specific functions, and have data objects (DO) qualified by data attributes (DA) and functional constraints (FC) that categorise their purpose [4]. The specification defines different ways to express IEC 61850 tags. In the context of this paper, we use the following:

$$\mathrm{Ph\_LD}\backslash\mathrm{DL.DO.DA.FC}$$

For example tag `PARK_ECP001_S3_SHL001_Inverter01\s4dinv.heatsinktmp.mag.f` can be decomposed as follows: `PARK` is the Ph and represents a PV park. `ECP001_S3_SHL001_Inverter01` is the LD and represents an inverter. `s4dinv` is the LN and represents the specific function of averaging some property of the inverter over a period of 10 minutes. `heatsinktmp` means that the property of interest is the heat sink temperature of the inverter. The DA.FC `mag.f` means magnitude expressed as a float.

These tags will be the headers of the CSV files which are extracted each week encompassing. Each file encompasses 250 tags. The aim is to transform these CSV files to a KG conformant with the Omega-X pattern.[5] To address this, we implemented a semantic Extract Transform Load (ETL) pipeline (illustrated in Figure 1) that transforms CSV files into a unified and understandable representation using KGs. Since the ontologies and data schemas were pre-defined in the project, and that the input data followed standard protocols, we adopted a top-down approach for KG creation [5]. The proposed pipeline starts after the extraction of the CSV files.

2.1. Pipeline tools

In the process of building KGs from heterogeneous data sources, RML (RDF Mapping Language) is used to define declarative rules that map data from formats like CSV, JSON, or XML, into RDF [6, 7]. However, manually writing RML rules for each dataset can be time-consuming and error-prone, especially when dealing with frequent updates or structurally similar files from different sources. To streamline this process, we leverage Jinja, a powerful Python templating engine, to dynamically generate RML rules. When a new dataset needs to be integrated, these templates are automatically populated with the specific configuration, producing valid RML mappings on the fly. To process RML files, we selected SDM-RDFizer [8] due to its open-source nature, ease of use, and Python-based implementation. For

[2] https://w3id.org/omega-x/repository

[3] Detailed information about the pilot site cannot be disclosed in this paper for confidentiality reasons.

[4] Substations are key facilities of the power grid where voltage is transformed and electricity is routed between generation, transmission, and distribution networks.

[5] The Omega-X pattern can be found on GitHub - https://github.com/NaveenVarmaK/Pipeline_Mapping_IEC61850_OmegaX-CSDM/images/pattern.jpg

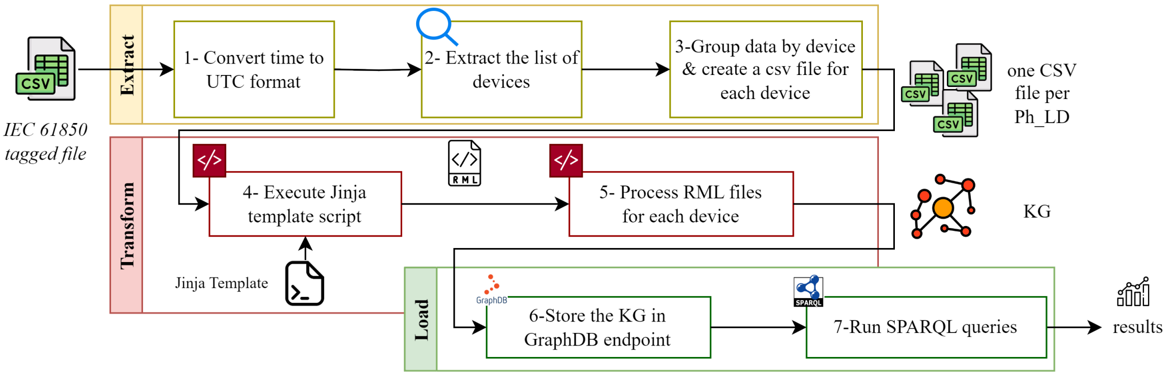
IEC 61850 tagged file
Extract
1- Convert time to UTC format
2- Extract the list of devices
3-Group data by device & create a csv file for each device
one CSV file per Ph_LD
Transform
4- Execute Jinja template script
Jinja Template
5- Process RML files for each device
KG
Load
6-Store the KG in GraphDB endpoint
7-Run SPARQL queries
results

Figure 1: ETL workflow for transforming IEC 61850 tag datasets into the OMEGA-X KG.

KG storage, Ontotext GraphDB [6] is used. as a triple store solution, tailored for storing, managing, and querying RDF data. It supports advanced features like reasoning and querying using SPARQL query language, which enables flexible and expressive querying over linked datasets.

2.2. Pipeline workflow

The pipeline is FAIRly available on GitHub[7], illustrated in 1 and consists of the following steps:

1. Extract: The extraction step begins with the normalisation of timestamps, converting all temporal data into the ISO 8601 date & time format with the UTC timezone. Following this, the list of devices is identified, and the dataset is segmented by device. For each device, a dedicated CSV file is generated. This step allows to attach the Ph_LD evaluation point to the data collection.
2. Transform: This stage constitutes the core of our semantic enrichment pipeline. The CSV files are transformed into KGs using RML mapping generated from a Jinja Template [8]. The RML files are then processed using SDM-Rdfizer to create the KG.
3. Load: Once generated, the KGs are loaded to a GraphDB triple store, which serves as the semantic data endpoint. The stored data is queried using SPARQL, enabling the validation and evaluation of the system against a set of competency questions.

To illustrate the proposed pipeline, we use the example of two CSV datasets, one containing meteorological data and another with inverter data from EDF pilot site. We first extract the list of devices from the CSV files to create the topology KG. Typically, the topology creation step is performed once and reused by the service providers to enable intelligent services. Weather stations and inverter datasets are then transmitted periodically, typically on a weekly basis. In order to determine the unit and the property, a rule-based parsing (regex + dictionary lookup) is being used. Regular expressions allow flexible pattern matching on text, while dictionary lookup ensures correct identification of known units and properties, improving accuracy and maintainability of the parsing system [9]. The dictionary contains for each LN and DO that may be found in the IEC 61850 tag, the corresponding unit from the QUDT unit vocabulary [9] and property from Omega-X property module. The produced KGs are stored in GraphDB end point. Figure 2 depicts a KG for `PARK_ECP001_S3_SHL001Inverter01`.

2.3. Pipeline Evaluation and lessons learned

To validate our pipeline, we conducted testing using data from **sixteen** inverters, **sixteen** weather stations, and approximately one **hundred** combiner boxes over **eighteen** weeks. Each week, the system processed around *1009* data points. The inverter data included DC current, DC voltage, DC power,

[6] https://www.ontotext.com/products/graphdb/
[7] All source files are available on GitHub - https://github.com/NaveenVarmaK/Pipeline_Mapping_IEC61850_OmegaX-CSDM
[8] https://jinja.palletsprojects.com/en/stable/templates/
[9] https://qudt.org/3.1.4/vocab/unit

```
@prefix ets: <https://w3id.org/omega-x/ontology/EventTimeSeries/> .
@prefix eds: <https://w3id.org/omega-x/ontology/EnergyDataSet/> .
@prefix prop: <https://w3id.org/omega-x/ontology/Property/> .
@prefix unit: <http://qudt.org/vocab/unit/>.
<PARK_ECP001_S3_SHL001Inverter01/week-1> a ets:DataCollection, eds:EnergyDataSet ;
  eds:includesEvaluationPoint <PARK_ECP001_S3_SHL001Inverter01> ;
  eds:isExchangedIn <PARK_ECP001_S3_SHL001Inverter01/week-1/EC> ;
  ets:comprises <PARK_ECP001_S3_SHL001Inverter01/week-1/4dinv.heatsinktmp> .
<PARK_ECP001_S3_SHL001Inverter01/week-1/4dinv.heatsinktmp> a ets:DataCollection ;
  ets:isAboutProperty prop:HeatSinkTemperature ;
  ets:hasUnit unit:DEG_C .
<PARK_ECP001_S3_SHL001Inverter01/week-1/4dinv.heatsinktmp_DP_6163972> a ets:DataPoint ;
   ets:belongsTo <PARK_ECP001_S3_SHL001Inverter01/week-1/4dinv.heatsinktmp> ;
   ets:hasPropertyValue
   ↪ <PARK_ECP001_S3_SHL001Inverter01/week-1/s4dinv.heatsinktmp_DP_6163972_PV6163972> ;
   ets:dataTime "2025-01-06 00:13:20"^^xsd:dateTime .
```

Figure 2: Portion of the KG showing the data collection and data points for the heat sink temperature of inverter `PARK_ECP001_S3_SHL001Inverter01`

enclosure temperature, heat sink temperature, and total active power. From the weather stations, we captured ambient temperature, plane of array insolation, and back-of-panel temperature. The evaluation of the produced KG focuses on its completeness, consistency, size and usability for intelligent services in renewable energy use cases. To assess the ***completeness*** of the KG, we executed a list of ***fifty*** CQs which are written as SPARQL queries and executed in GraphDB. Basically, the queries aim to retrieve the device, the property, unit of measure, timestamp, and the values. In addition, for predictive maintenance services, we used queries depicting the topology of the park. These queries retrieve the connected devices as well as some datasheet properties like maximum DC current and maximum DC voltage for each inverter. All needed information was successfully retrieved. The responses to these queries proved highly valuable for the service providers of the Omega-X project, who previously found it difficult to interpret the raw CSV files before semantic enrichment. To ensure semantic consistency, we used the Pellet reasoner [10]. Furthermore, we evaluated the ***number of triples*** of the produced KG. The weekly inverter KG consists of 10240 triples generated from a CSV file with 39 columns and 1009 rows, while the weekly weather station KG consists of 6144 triples generated from a CSV file with 19 columns and 1009 rows. The semantic enrichment significantly improves the data's utility and interoperability with other data from other data providers who were not using the IEC 61850 tags but reused the same Omega-X pattern to describe semantically their measurements. For example, the predictive maintenance service is tested on two distinct datasets provided by EDF (that uses IEC 61850) and ESTABANELL-EYPESA. This demonstration illustrates the potential of deploying services across different companies adopting heterogenous dataset format by relying on a common data representation model. Moreover, the shared taxonomy of properties used in the KG allows for the reuse of the dataset in flexibility services. To evaluate the usability of the produced KGs, we assessed their adoption by service providers. Four out of five service providers in Omega-X were able to successfully utilise the KGs to generate their services, demonstrating the practicality and effectiveness of the graph structure and data model. However, the remaining one encountered difficulties due to the absence of some inverter properties description, whose inclusion requires significant manual effort and time. On average, the complete pipeline for generating a weekly KG for a single device takes less than 1 minute. To better understand performance at a finer granularity, we analysed a representative execution using a CSV file for an inverter. All experiments and performance evaluations were conducted on a personal laptop running Ubuntu, equipped with the following hardware specifications: **CPU:** 11th Gen Intel® Core$^{\text{TM}}$ i7-1185G7 @ 3.00GHz, 4 cores, 8 threads; **CPU Frequency:** 400 MHz (min) – 4.8 GHz (max); **RAM:** 16 GB; **Operating System:** Ubuntu 22.04 LTS (64-bit). Over the span of 18 weeks, on average, the creation of the KG is done in 47.3 seconds and the peak memory usage is 75.20 MB.

3. Conclusion

The paper describes the steps of the automatic ETL pipeline for KG creation from IEC 61850 tags and structured using the Omega-X ontology. The proposed solution consists of three main components: Extraction of devices and grouping data per device; KG generation using Jinja and RML files; KG storage in GraphDB triple store for information retrieval using SPARQL queries defined from CQs. To the best of our knowledge, using text-based templates to generate RML mappings and then use these mappings automatically is an innovative approach that proved to be both simple and a pragmatic fit to our needs. RML mappings being expressed in RDF, an alternative could consist in using RML itself to generate RML mappings, although we believe it would be less understandable and maintainable. The produced KGs have been utilised by various service providers who consumed the KGs in the Omega-X project to deliver services related to predictive maintenance and energy prediction. This has improved the understandability and reusability of the datasets. Moreover, the application of semantic techniques has proven effective in addressing the interoperability challenges between datasets from the same data provider as well as across different data providers. Fortunately, the creation of these KGs paves the way for their application across diverse use case families. Notably, they enhance scenarios where energy production prediction is required, such as in flexibility related use cases. EDF is planning to test this pipeline in other departments than solar production where IEC 61850 is used for data exchange. However, a major challenge lies in the scale of these KGs, the diversity of data sources, and the complexity introduced by IEC 61850 tags. There is a clear need for adaptable approach where the mapping process should be generated based on the IEC 61850 tags encountered in the data.

4. Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] ENERShare Consortium, Blueprint of the Common European Energy Data Space (CEEDS), Technical Report, 2024. Accessed: 2025-05-07.

[2] OMEGA-X Consortium, D3.4: Data Analytic Services and Requirements Related to Interoperability, Security, Privacy, and Data Sovereignty, Technical Report, 2024. Accessed: 2025-05-07.

[3] Tatsoft, Iec 61850 overview and implementation, 2021. URL: https://tatsoft.com/wp-content/uploads/2021/10/IEC61850.pdf, accessed: 2025-04-01.

[4] E. Tebekaemi, D. Wijesekera, Designing an iec 61850 based power distribution substation simulation/emulation testbed for cyber-physical security studies, in: Proceedings of the First International Conference on Cyber-Technologies and Cyber-Systems, 2016, pp. 41–49.

[5] Z. Zhao, S.-K. Han, I.-M. So, Architecture of knowledge graph construction techniques, International Journal of Pure and Applied Mathematics 118 (2018) 1869–1883. URL: https://acadpubl.eu/jsi/2018-118-19/articles/19b/24.pdf.

[6] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, R. Van de Walle, Rml: A generic language for integrated rdf mappings of heterogeneous data., Ldow 1184 (2014).

[7] A. Iglesias-Molina, D. Van Assche, J. Arenas-Guerrero, B. De Meester, C. Debruyne, S. Jozashoori, P. Maria, F. Michel, D. Chaves-Fraga, A. Dimou, The rml ontology: a community-driven modular redesign after a decade of experience in mapping heterogeneous data to rdf, in: International Semantic Web Conference, Springer, 2023, pp. 152–175.

[8] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana, M.-E. Vidal, Sdm-rdfizer: An rml interpreter for the efficient creation of rdf knowledge graphs, in: Proceedings of the 29th ACM international conference on Information & Knowledge Management, 2020, pp. 3039–3046.

[9] L. Chiticariu, Y. Li, F. R. Reiss, Rule-based information extraction is dead! long live rule-based

information extraction systems!, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 827–832.

[10] S. Abburu, A survey on ontology reasoners and comparison, International Journal of Computer Applications 57 (2012).