Linking the Grid: A Knowledge Graph Approach to France's Electricity Consumption

Thibault Ehrhart[0000−0003−1377−8279] 1, Pasquale Lisena[0000−0003−3094−5585] 1, Raphael Troncy[0000−0003−0457−1436] 1, Ghislain Agoua[0000−0002−6072−7477] 2, Somsakun Maneerat[0000−0001−5745−0879] 2, and Fatma-Zohra Hannou[0000−0003−4747−1232] 2

[1] EURECOM, Sophia Antipolis, France
{thibault.ehrhart,pasquale.lisena,raphael.troncy}@eurecom.fr
[2] EDF R&D, SEQUOIA, Lab Paris-Saclay, Palaiseau, France
{somsakun.maneerat,ghislain.agoua,fatma-zohra.hannou}@edf.fr

Abstract. Understanding and monitoring electricity consumption is critical for ensuring the stability, efficiency, and sustainability of energy systems, particularly in the context of national energy transitions. In France, where electricity production is increasingly shaped by variable renewable sources and evolving consumption patterns, accurate and timely insights into demand dynamics are essential for both grid management and policy-making. However, electricity usage data is inherently heterogeneous, spanning diverse spatial scales and temporal resolutions. It originates from various sources and can be combined with other relevant data such as weather data and socio-economic indicators. In this work, we present our methodology for building a very large knowledge graph reflecting the electricity consumption in France enriched with numerous other data sources. We describe the ontologies we developed and re-used, the ETL conversion process and how we align the data on a fine-grained spatial and temporal scale. Next, we experiment with various graph embeddings techniques in order to compute the similarity between small geographical areas according to their electricity consumption pattern. We demonstrate that the symbolic knowledge injected in the knowledge graph enables us to obtain different similarities depending on macro criteria. Finally, we present an end-user web application that visualizes this very large knowledge graph while being very reactive thanks to various optimizations.

Keywords: Energy · Knowledge Graph · Data Integration · Graph Embeddings.

1 Introduction

The French energy system is undergoing a profound transformation driven by the national energy transition and the increasing share of variable renewable energy sources involved. This evolution has resulted in increasingly dynamic and localized electricity consumption patterns, thus challenging grid stability, forecasting of demand, and planning of infrastructure. This underlines the growing need to accurately understand detailed electricity consumption patterns at fine-grained spatial and temporal levels.

We observe that the information ecosystem in this field is very fragmented. In France, the energy consumption data for households, SMEs and large enterprises are released by local electricity distribution companies such as ENEDIS and aggregated in the open data portal of the Ministry of Ecological Transition, but also by the electricity transmission system operator (RTE), using different formats and spatio-temporal resolutions. At the same time, many other contextual datasets that have an impact on electricity consumption such as weather data, socio-economic indicators, and territorial infrastructure are available as open data sources from institutions such as Météo-France, INSEE, and IGN. However, these valuable datasets are often left separately, which in turn restrict their capabilities for integrated analysis. Having no single representation is a problem for making cross-domain queries, inferring causal relationships, or finding interesting similarity patterns in different territories.

In this paper, we tackle these challenges arguing that semantic web technologies are ideally positioned to realize this large scale data integration. This results in an RDF-based knowledge graph that integrates electricity consumption data at a fine-grained geographical resolution named IRIS[3] at daily, monthly and annual temporal granularity, and enriched with weather, demographic, housing characteristics, public equipment and other economic indicators. We show how this knowledge graph can be used to derive novel insights regarding the evolution of electricity consumption in France.
Our main contributions are as follows:

- We release a lightweight ontology that connects and selectively extends well-established ontologies, including SOSA/SSN [6], GeoSPARQL [11], Schema.org, and vocabularies released by the French Institute of Statistics and Economic Studies (INSEE)[4] and the Institute of Geographic and Forest Information (IGN)[5];
- We develop a modular Python-based ETL pipeline that converts multiple datasets into RDF with stable URIs and perform spatio-temporal alignments;
- We conduct an experimental evaluation using graph embeddings (RDF2Vec and engineered feature vectors) to infer fine-grained similarity between IRIS areas.
- We implement an interactive web application which provides a map-based visualization and exploration of the underlying knowledge graph data.

2 Related Work

Semantic modeling and knowledge graph construction for energy data have been the focus of several recent research efforts. Chun et al. [3] introduced the Energy

[3] An IRIS (Îlots Regroupés pour l'Information Statistique) is a micro neighborhood that has an average of 2,500 inhabitants.
[4] Ontologies developed by INSEE: `https://github.com/InseeFr/Ontologies`
[5] Ontologies developed by IGN: `https://data.ign.fr/data.html`

Knowledge Graph (EKG) with an ontology designed to integrate heterogeneous energy-related resources and support smart energy services in decentralized grids. In the open data domain, Hanžel et al. [5] presented a multi-region electricity consumption dataset and an RDF knowledge graph contextualizing household-level data with semantic annotations. Their work addresses challenges similar to ours by developing a uniform format for electricity consumption data across multiple regions. Their approach demonstrates the value of semantically enriched datasets for large-scale analysis and enables interoperability with open knowledge bases like Wikidata and DBpedia. Other initiatives, such as the Ontotext's Transparency Energy Knowledge Graph (TEKG) have focused on integrating pan-European electricity data (e.g. ENTSO-E) into RDF, showcasing the benefits of Linked Open Data for analytics and policy-making. While prior works focus on regional datasets or specific domains, our contribution uniquely targets a large-scale, nationwide knowledge graph for France which integrates diverse open datasets at multiple temporal and spatial scales. Hence, our knowledge graph better contextualizes the raw electricity consumption in integrating numerous other sources such as weather and socio-economic indicators.

Graph embedding techniques have been explored for knowledge graph analysis. Ristoski et al. [13] proposed RDF2vec, a method for generating node embeddings from RDF graphs using random walks and word embedding models. Extensions such as entity2rec [12] based on node2vec [4] have shown that embedding approaches can effectively capture both structural and semantic patterns in graph data. In our work, we emphasize both symbolic reasoning (ontology-based integration) and neural approaches (graph embeddings) to enable advanced analysis of electricity consumption patterns across French territories.

3 Data Sources and Spatio-Temporal Inference

The integration of electricity consumption and contextual data at fine-grained spatio-temporal granularity is challenging due to the heterogeneity of available sources, which vary considerably in terms of spatial resolution, geographic coverage, and update frequency.

3.1 Data Sources Overview

Table 1 summarizes the different datasets incorporated into the knowledge graph, indicating their spatial coverage and temporal resolution.

INSEE COG. The COG (*Code Officiel Géographique*) defines the administrative divisions of France. We use the original dataset published by INSEE in RDF, which includes an administrative decomposition of the country into regions, departments, districts (named *arrondissements*), municipalities (named *communes*), and finally IRIS zones, together with their containment relationships and historical changes. Each administrative unit is described by a persistent identifier, a name, a type[6],

[6] The different types of IRIS are described at `https://www.insee.fr/fr/information/2438155`

and a validity period. Every year, a number of changes can be made such as the division or merger of IRIS zones or municipalities, the modification of existing IRIS geo shapes, etc.

IRIS Contours. The IRIS contours dataset provides the precise geographical boundaries (polygons) for all IRIS units as shapefiles in the `EPSG:4326 (WGS 84)` coordinate format. Each IRIS polygon includes both a unique identifier and the coordinates (latitude/longitude) of the points. We load the shapefiles into `geopandas` dataframes in order to compute the neighbors.

ENEDIS. The ENEDIS dataset is further decomposed into two main datasets. First, it offers the annual electricity consumption statistics (from 2011 to 2023) at the IRIS geographical granularity. Each record includes the yearly total consumption (in MWh) divided by sector, where the sectors are: residential (households), professional (typically, local businesses), and industrial (typically, SMEs and large enterprises). Second, it provides a detailed description of the households park including the total number of housing units, whether they serve as primary or secondary housing, their construction period divided into seven temporal spans, their average size dividing into six levels, the proportion of collective housing, the share of electrical heating, etc.

RTE éCO2mix. The éCO2mix dataset is published by the French transmission system operator RTE and contains 15-minute interval electricity consumption measurements (actual, consolidated, and final) that cover France as a whole, its 12 administrative regions[7], and 21 large metropolitan areas. Each record includes the area, date, time, total consumption (in MW) and detailed breakdowns by energy source.

SDES. The SDES (*Service des Données et Études Statistiques*) publishes several energy related datasets. We use the dataset "Données de consommation et de points de livraison d'énergie à la maille IRIS - chaleur et froid" which provides the annual energy consumption (in MWh) and the number of delivery points for electricity, natural gas, heat, and cooling at IRIS level since 2018. The data is broken down by energy type and by five main economic sectors (agriculture, industry, tertiary, residential, and unassigned), and by two-digits NAF code when available.

Météo-France. The Météo-France dataset contains daily meteorological observations from 2,439 weather stations distributed nationwide. Each station has a fixed location (latitude/longitude), an identifier, and metadata such as the station type and altitude. For each day, we have collected three measures: the air temperature (°C), the wind speed (m/s), and the precipitation (mm).

[7] All mainland France regions, excluding Corsica.

Table 1. Overview of data sources integrated into the knowledge graph. Most datasets are static or aggregated at the IRIS level unless noted otherwise.

Source	Description	Temporal	Spatial
INSEE COG[a]	Administrative codes and geographical units.	static	municipalities, departments, regions
IGN IRIS[b]	Boundaries of IRIS zones.	static	IRIS
IGN AdminExpress[c]	Administrative boundary ontology.	static	municipalities, departments, regions
ENEDIS[d]	yearly (2011–2023) electricity use by sector.	yearly	IRIS
SDES[e]	Aggregated (2018–2022) from 111 operators.	yearly	IRIS
RTE éCO2mix[f]	Real-time (15-min) regional consumption.	15 min	metropolis, regions
RTE éCO2mix[f2]	Consolidated regional consumption (2013–2023).	30 min	metropolis, regions
Météo-France[g]	Daily weather (2023–2024) from 2,439 stations.	daily	weather stations
INSEE BPE[h]	Snapshot (2024-05-09): 2M+ facility locations.	static	IRIS
INSEE SIRENE[i]	Snapshot (2024-05-02): 15M+ organizations (NAF).	static	IRIS

[a] https://rdf.insee.fr/geo/index.html
[b] https://www.data.gouv.fr/en/datasets/contours-iris-r/
[c] https://www.data.gouv.fr/fr/datasets/admin-express/
[d] https://data.enedis.fr/explore/dataset/consommation-electrique-par-secteur-dactivite-iris/information/?sort=annee
[e] https://www.statistiques.developpement-durable.gouv.fr/catalogue?page=dataset&datasetId=610244c9e436671e84ec5da8
[f] https://odre.opendatasoft.com/explore/dataset/eco2mix-regional-tr/information/?disjunctive.nature&disjunctive.libelle_region
[f2] https://odre.opendatasoft.com/explore/dataset/eco2mix-regional-cons-def/information/?disjunctive.nature&disjunctive.libelle_region
[g] https://meteo.data.gouv.fr/datasets/6569b51ae64326786e4e8e1a
[h] https://www.data.gouv.fr/fr/datasets/base-permanente-des-equipements-1/
[i] https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/

BPE (Base Permanente des Équipements). This dataset includes all public facilities and amenities (over 2 million units), ranging from schools and health centers to retail and leisure facilities. The BPE relies on a controlled vocabulary that aims to categorize the facilities. We have developed and published a Python script[8] that generates a controlled vocabulary[9] in RDF where each of the 222 facility types is described as a SKOS concept forming a three level hierarchy.

SIRENE. The SIRENE database is the official business directory of France. It lists all organizations and their establishments and contains more than 15 million entries. Each establishment includes a SIRET identifier, precise location, legal status, workforce size, and its principal activity code following the NAF (*Nomenclature d'Activités Française*). We used the RDF vocabulary published by INSEE[10], which provides a SKOS hierarchy for the 1,733 economic activities.

3.2 Spatio-Temporal Gaps and Inference

As shown in the Table 1, the data sources vary along two main dimensions:

– **High temporal, low spatial resolution:** the RTE éCO2mix dataset provides a near-live observation of the electrical consumption in France

[8] https://github.com/D2KLab/bpe-vocabulaire
[9] https://github.com/D2KLab/bpe-vocabulaire/releases/download/v0.2.0/bpe-vocabulaire.ttl
[10] https://rdf.insee.fr/codes/index.html

(sampled every 15 minutes), but only at a coarse-grained geographic level (covering regions and metropolitan areas), which is insufficient for IRIS-based analysis. For example, *Région Sud*, one of the 13 regions in France, is composed of 2,434 IRIS.
- **High spatial, low temporal resolution:** the ENEDIS and SDES datasets provide the electricity consumption at the IRIS-level. However, they only provide the annual electricity consumption per sector.

Consequently, we standardize all contextual and consumption data at the IRIS spatial granularity and daily or monthly temporal granularity through the inference and alignment pipeline described below. The resulting knowledge graph targets retrospective territorial analysis rather than real-time grid monitoring, since the IRIS-level consumption data have annual granularity and are strongly influenced by seasonal patterns.

Temporal Downscaling. We transform yearly IRIS consumption data (from ENEDIS and SDES) into daily and monthly time series using standardized profile coefficients (*coefficients de profils*) constructed by ENEDIS and validated by the Commission de Régulation de l'Énergie[11]. These profiles are built from representative site data and provide half-hourly dynamic coefficients for different customer categories (residential, professional, industrial). The profiles exist in two distinct forms: *static profiles* use predefined average coefficients adjusted for reference or actual weather conditions, while *dynamic profiles* provide more precise half-hourly values based on measured power from the sample panel.

We aggregate these half-hourly profiles to compute average weights at the desired temporal resolution (monthly, weekly, or daily). For each annual electricity consumption record associated with an IRIS zone, we perform temporal disaggregation by applying the corresponding aggregated profile coefficients as follows:

1. **Profile Selection:** Based on the category (e.g. residential, professional, industrial), we select the appropriate profile.
2. **Weight Application:** For each category, the annual consumption value is proportionally distributed according to the profile weights. To ensure that the sum matches the original annual total, we compute a correction factor K_{corr} as the ratio between the recorded annual consumption and the sum of the selected profile's weights:

$$K_{\text{corr}} = \frac{\text{Annual\_Consumption}}{\sum_{i=1}^{n} \text{Profile\_Coefficient}_i}$$

where n is the number of periods (e.g. $n = 12$ for months, $n = 365$ for days). Each period's disaggregated consumption is then calculated as:

$$\text{Consumption}_i = K_{\text{corr}} \times \text{Profile\_Coefficient}_i$$

[11] https://data.enedis.fr/explore/dataset/coefficients-des-profils/information/

This disaggregation process yields monthly and daily consumption estimates for each IRIS zone. We validated our approach by comparing the aggregated results against éCO2mix's 15-minute consumption data for the *Région Sud*. The disaggregated series closely matched the observed trends, confirming that the standardized profile coefficients provide a reliable temporal breakdown from the annual aggregates (Figure 1).

Spatial Projection. For datasets with precise geo-locations (e.g. weather stations, SIRENE establishments, BPE facilities), we assign those locations to IRIS polygons using spatial joins. To handle millions of records in SIRENE and BPE, we built an R-tree[12] spatial index and used it to rapidly determine the IRIS membership of each establishment or facility point.

Aggregation. For datasets like SIRENE and BPE, we perform aggregation to summarize data at the IRIS level. We are not interested in the detailed listing that they provide, but instead, our goal is to describe each IRIS according to the number of establishments and facilities they contain. Specifically, we group SIRENE establishments by IRIS and NAF code and compute counts and employee size distribution. We do the same with BPE by grouping the facilities by first-level category.

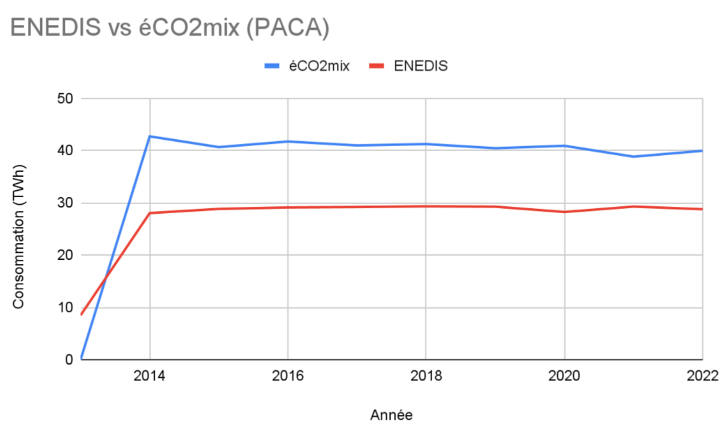

Fig. 1. Electrical consumption in TWh in the PACA region in 2014-2022 according to the real éCO2mix data and the daily inferred data from ENEDIS using the profile coefficients. There is a stable ∼10 TWh difference since major electricity consumers (e.g. the national railway operator) are direct clients of RTE and do not go through distribution providers such as ENEDIS.

4 Ontology and Knowledge Graph Population

In this section, we detail the ontologies that have been used to integrate the datasets listed in Table 1 as well as our knowledge graph construction workflow.

[12] `https://pypi.org/project/Rtree/`

4.1 Ontology Design

Our knowledge graph design follows a hybrid approach by re-using existing ontologies from Schema.org, SKOS, SOSA, GeoSPARQL, and INSEE vocabularies, while introducing additional classes and properties for domain-specific concepts. The ontology is published at `http://data.edf.eurecom.fr/ontology/`.

The ontology primarily acts as an integration layer between vocabularies selected for their relevance, adoption, and compatibility with French public-administration datasets. New terms were introduced only where the reused ontologies did not provide adequate coverage, particularly for describing the housing stock of a territory; in these cases, the properties reflect the concepts represented by the original dataset columns. The contribution therefore lies mainly in the integrated modeling and operational use of these heterogeneous data sources, consistently with the application-oriented scope of this work.

IRIS Zones. The central entity in our knowledge graph is the French IRIS zone. Each IRIS is represented by a base URI in the form `{base_uri}/iris/{iris_code}`. To capture yearly contextual information associated with an IRIS (e.g. population, housing statistics from the ENEDIS dataset), we create yearly instances typed as `insee:TimeSeries`, `dcat:Dataset`, and `prov:Entity` following the pattern `{base_uri}/iris/{iris_code}/{year}`. These yearly instances are linked back to the base IRIS URI via `schema:description` and hold properties such as `schema:startDate`, `schema:endDate`, `insee:validFrom`, `insee:validUntil`, along with other statistical attributes mapped from the source data (e.g. `edf:tauxLogementsCollectifs`, `idemo:populationTotale`).

Observation Pattern. For representing time-varying data points (electricity consumption, weather measurements, aggregated counts), we use `schema:Observation` either directly or as a subclass. For meteorological data, we leverage the SOSA ontology by using `sosa:Observation` alongside `schema:Observation`. As illustrated in Figure 2, observations are distinct resources, typically identified by UUID-based URIs derived deterministically from key attributes (see Section 4.2). Key properties associated with observations include:

- `schema:observationAbout` or `sosa:featureOfInterest`: Linking the observation to the entity it describes (typically an IRIS zone, a yearly IRIS instance, or a weather station located within an IRIS).
- `schema:observationDate` or `sosa:resultTime`: Recording the timestamp or date of the observation (using `xsd:date`, `xsd:gYearMonth`, or `xsd:gYear` depending on the granularity).
- `schema:value` or `sosa:hasSimpleResult`: Holding the measured value as a literal with an appropriate XSD datatype, often including units from QUDT.
- `schema:variableMeasured` (for aggregated statistics) or `sosa:observedProperty` (for sensor readings): Linking to a concept representing what was measured.

Electricity Consumption. Annual consumption records are represented using the custom class `edf:Consommation`. These instances link to the yearly IRIS TimeSeries via `schema:observationAbout` and include properties like `edf:codeCategorieConso`, `edf:codeGrandSecteur`, and the consumption value (`schema:value`) typed with `unit:MegaW-HR` from the QUDT namespace. Our inferred monthly and daily consumption estimates generated via temporal disaggregation (see Section 3.2) also follow this pattern, using appropriate `xsd:gYearMonth` or `xsd:date` literals for `schema:observationDate`.

Weather Data. Following SOSA best practices, weather stations are modeled as `sosa:Platform` and `schema:Place`, with properties like `schema:name`, `schema:latitude`, and `schema:longitude`. Stations are linked to the IRIS zone they fall within using `schema:containedInPlace`. Individual sensing devices, such as temperature, wind, and precipitation sensors, are represented as `sosa:Sensor` instances and associated with their corresponding weather stations. Materialized `sosa:Observation` instances capture daily readings (`sosa:hasSimpleResult` typed with appropriate units like `unit:DEG_C` or `unit:M-PER-SEC`) and link back to the station (`sosa:madeBySensor` implicitly linked via the sensor URI structure) and the property observed (`sosa:observedProperty`).

Socio-Economic and Infrastructure Data (SIRENE, BPE). Aggregated counts from SIRENE (establishments per NAF category) and BPE (facilities per type) at the IRIS level are modeled using `schema:Observation`. To define the nature of these counts, we introduce `schema:StatisticalVariable` instances. Each `schema:StatisticalVariable` follows the pattern `{base_uri}/sirene/{naf_category}` and represents a specific metric (like "Number of establishments in NAF category X"), links to the population type using external vocabulary URIs (e.g. INSEE NAF codes, BPE vocabulary URIs), specifies `schema:statType` as `schema:count`, and has a human-readable `schema:name`. The `schema:Observation` then links to the IRIS (`schema:observationAbout`), the date (`schema:observationDate`), the count value (`schema:value` typed as `xsd:integer`), and the corresponding `schema:StatisticalVariable` via `schema:variableMeasured`. Labels for NAF and BPE categories are incorporated using `skos:prefLabel`.

Figure 2 shows an excerpt of the knowledge graph for the IRIS 920730116, corresponding to a neighborhood of the Suresnes municipality.

4.2 Knowledge Graph Population

We populate the KG through a multi-step ETL process for each data source. Since the datasets differ considerably in format (CSV, JSON, shapefiles, APIs), temporal resolution (15-minute up to yearly readings), and spatial granularity (point coordinates, regions, IRIS polygons), we developed ad-hoc ETL components rather than RML mappings. Materialization was motivated by the extensive preprocessing required, including R-tree spatial joins, profile-based temporal downscaling, aggregation, normalization, and URI generation. Nevertheless, we

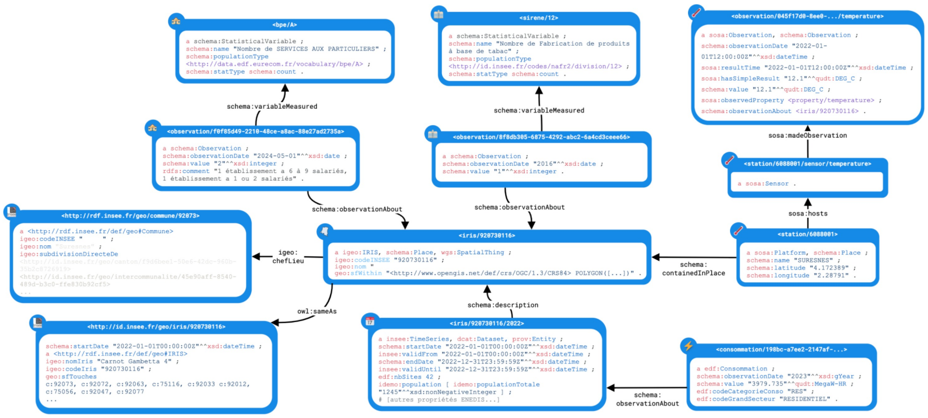

Fig. 2. Visualization of multiple observations linked to IRIS 920730116.

adhere to RDF best practices by re-using standard vocabularies and controlling graph generation through deterministic direct mappings.

For each source referenced in Table 1, we have written converters that read raw data, clean and normalize it, and output RDF triples according to the ontology described in Section 4.1. The code is available at `https://gitlab.eurecom.fr/edf-knowledge-graph/knowledge-graph`.

Extraction. We extract data from its native format using Python and the `pandas` library. Each source has its own processor class responsible for the extraction step, including reading files and applying basic preprocessing such as column renaming, filtering, and geocoding where necessary.

Transformation. In the transformation step, we convert the extracted data into RDF triples. We assign persistent and dereferenceable URIs to all consumption records. We construct these URIs deterministically by hashing the following key attributes using SHA-1: IRIS code, date, and category. For example, a consumption observation for a given IRIS and year would have a URI like `http://data.edf.eurecom.fr/consommation/{uuid}`, where `{uuid}` is derived from the concatenation of the relevant attributes.

Loading. Finally, we load the transformed data by generating RDF triples with `rdflib`. We first build the triples in an in-memory graph and periodically serialize them into N-Triples files to optimize performance. This approach is crucial given the data volume: daily consumption data alone amounts to approximately 45 GB (300 million triples) per year, and the full dataset covering 2011 to 2023 is estimated at over 600 GB (4 billion triples) for the electricity consumption alone. Table 2 presents the summary statistics of the data currently loaded in our knowledge base, which covers the whole territory of France for the year 2023. This snapshot is served by Virtuoso 7.2 on a 24-core Intel Xeon E5-2670 server with 256 GB of RAM, of which the loaded graph uses approximately 80 GB.

Table 2. Summary statistics of the knowledge graph data, France, year 2023.

Metric	**Count**
Total Triples	420,307,869
Instances of `schema:StatisticalVariable`	96
Instances of `edf:Consommation`	58,728,334
Instances of `insee:TimeSeries`	625,635
Instances of `igeo:IRIS`	49,542
Instances of `sosa:Sensor`	5,365
Instances of `sosa:Platform`	2,439
Instances of `schema:Observation`[a]	4,056,903

[a] Includes 2,310,417 entities that are both `sosa:Observation` and `schema:Observation`

We evaluated query-serving scalability using nine SPARQL queries employed by the EDF Conso Explorer, with 1, 8, 32, and 128 parallel clients. Median latency increased from approximately 10 ms for a single client to 354 ms with 128 clients, while the 99th-percentile latency remained below 750 ms and throughput plateaued at approximately 260 queries per second. These results indicate that the endpoint remains responsive under loads exceeding the expected application usage; application-level caching could further reduce the cost of repeated queries.

5 EDF Conso Explorer: Navigating Electricity Consumption Through Linked Data

In this section, we demonstrate two applications that make use of the knowledge graph: an experiment that aims to derive fine-grained similarity measures between two IRIS according to their raw electricity consumption and other factors (Section 5.1) and an end-user web application that enables users to efficiently search and visualize the knowledge graph despite its size (Section 5.2).

5.1 Graph Embeddings for Similarity Computation

Here, territorial similarity is investigated as one illustrative downstream application of the presented knowledge graph. ENEDIS provides services[13] that enable to get a summary of the electricity production and consumption pattern of a given municipality in order to advise on future policies for transitioning to higher renewal energy usage. One of the service is specifically dedicated to identifying similar territories for a given municipality.[14] However, this service takes into account only population and housing characteristics. We selected this task because territory comparison is of direct practical interest to EDF.

[13] `https://openservices.enedis.fr/bilan-de-mon-territoire/`
[14] `https://openservices.enedis.fr/territoires-similaires/`

Our goal is to demonstrate that the enriched KG described in Section 4 can support more fine-grained similarity measures incorporating macro-criteria such as weather and socio-economic activity. Our approach consists of embedding the KG and computing cosine similarities between the resulting vectors. In this experiment, we wanted to compare two different approaches producing graph embeddings (RDF2Vec and a handcrafted feature-vector approach) in order to demonstrate the added value of incorporating symbolic knowledge and capturing taxonomic reasoning via the hierarchical controlled vocabularies.

Data preprocessing. In order to produce meaningful embeddings, we applied some common preprocessing steps to the data retrieved from the knowledge graph and processed with RDFLib [8]. These steps include:

- exclude literal properties that do not provide learnable information, such as class and entity labels, identifiers, and textual comments;
- parse other literal values, in particular, the numerical ones and discretize them into buckets.

Embedding with RDF2Vec. Random-walk-based methods are particularly effective for approximating graph properties such as node proximity, similarity, and structural features of nodes and subgraphs. These approaches simulate the movement of a "walker" through the graph. At each step, the walker selects one of the current node's outgoing edges at random and traverses it to the next node. The result is a sequence of visited nodes, known as a "walk", which serves as the basis for further analysis. The generated walks can be viewed as analogous to sentences in a text corpus, enabling the application of word embedding techniques like *word2vec*. Methods such as *RDF2Vec* [13] or *entity2vec* [12] follow this approach. By treating nodes as "words" and walks as "sentences", these methods model semantic relationships between nodes. Nodes that frequently co-occur in the same walks are considered semantically close, which is reflected in their placement within the learned embedding space, where similar nodes are positioned near one another. This approach effectively captures both structural and semantic relationships within the graph.

We employ RDF2vec – in the PyRDF2vec implementation [14] – given its popularity and good performance across a range of tasks. RDF2vec is suitable for object properties that connect nodes of the graph, but it was not designed to handle literal values, especially numeric ones. For this reason, we apply a further preprocessing that aims to map all (potentially continuous) literal values to discrete steps. The range of values to be mapped is calculated independently for each data property, using a linear scale based on the minimum and maximum values within the graph's dataset. The number of steps is determined as the minimum of 10 and the number of distinct values in the graph. Consequently, we consider a graph that contains all object properties and all data properties, and the literal values of the latter have been replaced by a handcrafted URI, identifying a step in the discrete range of that property[15].

[15] An exception is the property `schema:value`, which is used in different contexts in the knowledge graph, so that separate ranges are created for each of them, using

Engineered Vectors. We implement a baseline strategy that does not generate graph embeddings, but instead concatenates the numerical values retrieved from the knowledge graph into explicit feature vectors without latent feature learning. First, we use the SPARQL endpoint for retrieving the list of all IRIS nodes. For each of these nodes, we retrieve all literal numeric values:

1. Directly linked to the IRIS node;
2. Linked to a *TimeSeries* entity connected to the IRIS through a `schema:description` property;
3. Linked to a *Consommation* entity connected to one of the *TimeSeries* mentioned in step 2 through a `schema:observationAbout` property;
4. Linked to a *Observation* entity connected to the IRIS through a `schema:description` property or through a `schema:observationAbout` property.

For steps 3 and 4, given that the number of values is variable and can be quite large, we decided to aggregate the values and only store the mean and the standard deviation. In both cases, it was possible to group the values by the categories indicated by `edf:codeCategorieConso` and `schema:variableMeasured` respectively for *Consommation* and *Observation*. In other words, we take into account the mean and the standard deviation for each category of those classes. All the extracted numeric values are concatenated so that, for each IRIS, we have an ordered sequence of numbers, with the missing values replaced by '0'. This set of vectors is then converted into the same format of RDF2vec in order to benefit from the same pipeline.

Evaluation. We intend to automatically replicate the identification of similar IRIS made on the previously mentioned service by ENEDIS, which provides the top three most similar neighborhoods, albeit through manual curation. Evaluating the retrieved similarities is challenging because no established ground truth defines which IRIS zones should be considered similar across the heterogeneous dimensions represented in the knowledge graph. We therefore adopt an indirect evaluation: for each seed IRIS, we assess whether the territories retrieved from the contextual representations exhibit similar monthly electricity consumption patterns. The resulting scores evaluate consistency with consumption behavior rather than providing a definitive judgment of territorial similarity.

A test set comprising 14 IRIS was selected by EDF, ensuring diversity in characteristics such as population (more and less densely populated areas), consumption patterns (high and low), and *temperature sensitivity* (how strongly electricity consumption responds to external temperatures). In this way, this test set is fairly representative of the different types of IRIS composing France.

Utilizing the embeddings, we identified, using each of the 14 IRIS as seed s, the three most similar target IRIS t_i within the embedding space by applying the Nearest Neighbor method on cosine distance. Then, the cosine distance is used again to compute the similarity between s and each t_i on their monthly

property chains. In addition, given the high number of values, we aggregate them and average the vector at IRIS level.

Table 3. Average similarity scores for the two methods (RDF2Vec and Engineered Vectors), trained on subgraphs including only population and housing characteristics.

Method	Trained on: **Population**			**Housing**		
	AVG	MAX	MIN	AVG	MAX	MIN
RDF2vec	0.59	0.99	0.06	0.40	0.63	0.21
Eng Vectors	0.50	0.98	0.07	0.51	0.78	0.27

consumption, obtained through the *Temporal Downscaling* method as explained in Section 3.2. In particular, a similarity score is computed as the complement of the normalised cosine distance D:

$$score(s, t_i) = 1 - \frac{D(t_i, s) - D_{min}}{D_{max}} \tag{1}$$

where D_{min} and D_{max} are the minimum and maximum distance between any two known IRIS in terms of monthly electricity consumption. The three resulting scores are then averaged.

Some initial results are presented in Table 3. Although the results do not clearly favor one of the two proposed representation methods, we can derive some valuable insights:

- the population density has a strong impact on the consumption estimation, making it possible to identify IRIS with up to 99% similar consumption. However, relying only on this information is not sufficient, as shown by the low minimum scores;
- housing information yields more stable results, but far from reaching optimal matches;
- improved results may be achieved not only by combining different types of information during the training phase but also by leveraging the strengths of each method. For example, an ensemble strategy can exploit complementary strengths of the two methods in handling different configurations.

5.2 EDF Conso Explorer

To make the knowledge graph accessible without requiring expertise in RDF or SPARQL, we developed the *EDF Conso Explorer*, an interactive web-based tool that provides a visual exploration of the electricity consumption in France. It has been designed as a generic visual interface for the knowledge graph to support multiple user profiles and use cases. Its primary users are energy-domain experts at EDF, who can exploit the similarity search to identify relevant correlations and gather input features for training AI models. The application is also designed to support analysts working with Distribution System Operators (DSOs), interested in investigating trends in specific IRIS zones, and comparing similar zones or consumption patterns based on socio-economic and environmental characteristics.

This can also help with decision-making and reporting activities. Territorial public services, local authorities, and individual consumers represent potential future user groups, with the goal of optimizing their public consumption schemes and offering relevant public services. Finally, the application is designed with a user-friendly and interactive interface that is suitable for a wider audience, e.g. to help consumers understand consumption trends and local patterns. This is particularly useful in a context where energy consumption and production are prominent societal issues. Moreover, extending the scope of users to reach a wider audience is part of the transition from internal company use to promoting its added value to the aforementioned potential end-users. The tool is publicly available at `https://conso-explorer.tools.eurecom.fr`.

System Architecture. The tool is implemented as a React-based single-page application (Next.js/TypeScript). Application data is retrieved from the SPARQL endpoint at query time using declarative queries made with *SPARQL Transformer* [10, 9] which enables direct transformation of SPARQL results into structured JSON-LD data. The results are then rendered into various charts using the VisX[16] charting library.

User Interface. The main interface displays a map of France segmented into 49,542 IRIS zones. Users can explore local electricity consumption and see how it correlates with factors such as housing characteristics and business activity. The tool also offers a similarity search feature that allows users to find territories similar to a selected IRIS zone, based on the embedding vectors described in Section 5.1. In particular, the tool offers four macro-level similarity measures:

- population: where the monthly (residential only) electricity consumption and the IRIS population are taken into account;
- weather: where the monthly (residential only) electricity consumption and the 3 weather-related observations (temperature, wind, precipitation) are taken into account;
- housing: where the monthly (residential only) electricity consumption and all the characteristics describing the housing (year of construction, size, usage, etc.) are taken into account;
- socio-economic: where the monthly professional and enterprise electricity consumption and the number and types of facilities and establishments (including their workforce size) are taken into account.

6 Conclusion and Future Work

We have presented the construction of a large-scale Knowledge Graph that integrates electricity consumption data across France at fine-grained spatial and temporal resolutions. We aligned heterogeneous datasets and modeled them

[16] `https://github.com/airbnb/visx`

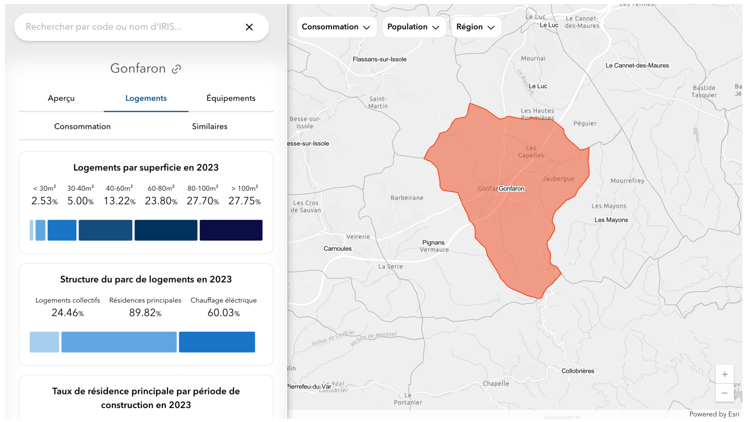

Fig. 3. Main interface of EDF Conso Explorer showing electricity consumption for a specific IRIS zone.

with a hybrid ontology combining standard vocabularies and domain-specific extensions. Our modular ETL pipeline efficiently transforms a large volume of data into RDF, with spatio-temporal alignments. Our initial experiments with graph embedding techniques highlight the potential of using symbolic knowledge to compute meaningful similarities between geographic areas, based on their consumption patterns and contextual factors. All developed tools have been released as open source, enabling other organizations to conduct similar studies.

Several directions can be explored to enhance and extend this work. The quality of the embeddings can be improved by normalizing engineered feature vectors (e.g. within the range $[-1, +1]$) and by selecting the most informative properties. Systematic ablation studies will also be conducted to quantify the contribution of each data source and feature type to similarity computation. Beyond territorial similarity, additional downstream tasks will be investigated, including the use of knowledge graph context to provide complementary explanations for the outputs of existing time-series energy forecasting models.

Future work will also include a task-based evaluation with EDF engineers and energy-domain researchers to assess the usefulness of fine-grained territorial similarity for activities such as IRIS clustering, correlation analysis, and feature selection. A complementary utility and usability study will involve potential end-users, including local authorities and individual consumers, to evaluate the accessibility and practical relevance of the application. We will also investigate the use of declarative RDF shape languages, such as SHACL [7] and ShEx [1], to specify and automatically validate structural, cardinality, datatype, and value constraints over the generated graph. Finally, we will study incremental refresh strategies [15] and knowledge graph virtualization [2] to reduce the cost of updating the materialized graph when new source data become available.

Supplemental Material Statement: The knowledge graph and the source code of all tools shown in this paper is thoroughly documented at `https://gitlab.eurecom.fr/edf-knowledge-graph`.

Declaration of use of Generative AI

During the preparation of this work, the authors used GPT5.2 in order to: grammar and spell check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

1. Boneva, I., Labra Gayo, J.E., Prud'hommeaux, E.G.: Semantics and Validation of Shapes Schemas for RDF. In: The Semantic Web – ISWC 2017. Lecture Notes in Computer Science, vol. 10587, pp. 104–120. Springer (2017). https://doi.org/10.1007/978-3-319-68288-4_7
2. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., Xiao, G.: Ontop: Answering SPARQL Queries over Relational Databases. Semantic Web **8**(3), 471–487 (2017). https://doi.org/10.3233/SW-160217
3. Chun, S., Jung, J., Jin, X., Seo, S., Lee, K.H.: Designing an integrated knowledge graph for smart energy services. The Journal of Supercomputing **76**(10), 8058–8085 (Oct 2020). https://doi.org/10.1007/s11227-018-2672-3
4. Grover, A., Leskovec, J.: node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 855–864. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939754
5. Hanžel, V., Bertalanič, B., Fortuna, C.: Towards data-driven electricity management: multi-region uniform data and knowledge graph. Scientific Data **12**(1) (Jan 2025). https://doi.org/10.1038/s41597-024-04310-z
6. Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M.: SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics **56**, 1–10 (May 2019). https://doi.org/10.1016/j.websem.2018.06.003
7. Knublauch, H., Kontokostas, D.: Shapes Constraint Language (SHACL). W3C recommendation, World Wide Web Consortium (Jul 2017), `https://www.w3.org/TR/shacl/`
8. Krech, D., Grimnes, G.A., Higgins, G., Hees, J., Aucamp, I., Lindström, N., Arndt, N., Sommer, A., Chuc, E., Herman, I., Nelson, A., McCusker, J., Gillespie, T., Kluyver, T., Ludwig, F., Champin, P.A., Watts, M., Holzer, U., Summers, E., Morriss, W., Winston, D., Perttula, D., Kovacevic, F., Chateauneu, R., Solbrig, H., Cogrel, B., Stuart, V.: RDFLib (Oct 2024). https://doi.org/10.5281/zenodo.6845245, `https://github.com/RDFLib/rdflib`
9. Lisena, P., Meroño-Peñuela, A., Kuhn, T., Troncy, R.: Easy Web API Development with SPARQL Transformer. In: 18th International Semantic Web Conference (ISWC), In-Use Track. Auckland, New Zealand (2019)
10. Lisena, P., Troncy, R.: Transforming the JSON Output of SPARQL Queries for Linked Data Clients. In: WWW'18 Companion: The 2018 Web Conference Companion. ACM, New York, NY, USA (2018). https://doi.org/10.1145/3184558.3188739

11. Matthew Perry, John Herring: OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Implementation Standard OGC 11-052r4, Open Geospatial Consortium (Sep 2012), `http://www.opengis.net/doc/IS/geosparql/1.0`
12. Palumbo, E., Monti, D., Rizzo, G., Troncy, R., Baralis, E.: entity2rec: Property-specific knowledge graph embeddings for item recommendation. Expert Systems with Applications **151**, 113235 (2020). https://doi.org/10.1016/j.eswa.2020.113235
13. Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., Paulheim, H.: RDF2Vec: RDF graph embeddings and their applications. Semantic Web **10**(4), 721–752 (2019)
14. Steenwinckel, B., Vandewiele, G., Agozzino, T., Ongenae, F.: pyRDF2Vec: A Python Implementation and Extension of RDF2Vec. In: European Semantic Web Conference (ESWC). pp. 471–483 (2023). https://doi.org/10.1007/978-3-031-33455-9_28
15. Vidal, V.M.P., Pequeno, V.M., Arruda Júnior, N.M., Casanova, M.A.: Publication and Maintenance of RDB2RDF Views Externally Materialized in Enterprise Knowledge Graphs. International Journal of Web Information Systems **18**(5/6), 255–285 (2022). https://doi.org/10.1108/IJWIS-02-2022-0043