Interlinking Heterogeneous Data for Smart Energy Systems

Fabrizio Orlandi[1], Alan Meehan[1], Murhaf Hossari[1], Soumyabrata Dev[1], Declan O'Sullivan[1], and Tarek AlSkaif[2]
[1]The ADAPT SFI Research Centre, Trinity College Dublin, Dublin, Ireland
[2] Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands

***Abstract*—Smart energy systems in general, and solar energy analysis in particular, have recently gained increasing interest. This is mainly due to stronger focus on smart energy saving solutions and recent developments in photovoltaic (PV) cells. Various data-driven and machine-learning frameworks are being proposed by the research community. However, these frameworks perform their analysis - and are designed on - specific, heterogeneous and isolated datasets, distributed across different sites and sources, making it hard to compare results and reproduce the analysis on similar data. We propose an approach based on Web (W3C) standards and Linked Data technologies for representing and converting PV and weather records into an Resource Description Framework (RDF) graph-based data format. This format, and the presented approach, is ideal in a data integration scenario where data needs to be converted into homogeneous form and different datasets could be interlinked for distributed analysis.**

NOMENCLATURE

API	Application Programming Interface
CSV	Comma-Separated Values
DB	Database
JSON	JavaScript Object Notation
KG	Knowledge Graph
PV	Photovoltaic
PWA	Photovoltaic and Weather Analysis
R2RML	RDB to RDF Mapping Language
RDB	Relational Database
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
SSN	Semantic Sensor Network Ontology
W3C	World Wide Web Consortium (Web standards and Linked Data technologies)

I. INTRODUCTION

With the recent developments in photovoltaic (PV) cells, there has been a renewed interest in the area of solar energy generation and forecasting. Most of the work in solar analytics involves mining data and proposing data-driven, machine-learning frameworks. These data are of diverse types, and are generated from various sensors [1]–[3]. Also such data comes in many forms (e.g. geological, medical, census, weather) and formats (e.g. RDB, CSV, JSON, APIs).

A. Motivation & Background

It is important to systematically analyse the diverse data types, such that we can identify trends, make predictions and inform decision-makers. Unfortunately, data is usually distributed across different sites and sources, and requires increasing amount of manual effort in discovering relevant pieces of information and pre-processing it. We can further unlock the potential of data if it is 'linked' and 'machine-readable'. The Web is moving away from information that is purely for human consumption, and instead expanding to include machine-readable data. This machine-readable data is being published in a format that allows computers to automatically understand how different pieces of data are connected to one another. This data[1] is known as *Linked Data* [4].

B. Relevant Literature

Linked Data follows a decade of research in the Semantic Web domain [5] and has recently reached considerable popularity under the name of 'knowledge graphs' (KGs) [6]. KGs have gained increasing popularity over the last years, especially in industry where they are now at the core of relevant consumer products (e.g. Google[2] and Bing[3] search engines). According to a recent business report [7], "51 percent of global data and analytics technology decision makers are either implementing, have already implemented, or are upgrading their graph database". KGs are knowledge-bases of facts about entities and concepts (e.g., places, persons, artifacts) which are represented using the flexible structure of a graph. Facts are often extracted from encyclopedic knowledge, such as Wikipedia, or existing structured repositories (e.g. Wikidata[4]), or even from unstructured sources such as social media posts (e.g. Facebook Graph[5]). More details of the relevant literature can be found in Section II.

This work is supported by the the Joint Programming Initiative (JPI) Urban Europe project: PARticipatory platform for sustainable ENergy managemenT (PARENT) and the Netherlands Science Foundation (NWO).

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Send correspondence to T. AlSkaif, E-mail: t.a.alskaif@uu.nl

978-1-7281-1156-8/19/$31.00

[1]https://lod-cloud.net/

[2]https://developers.google.com/knowledge-graph/

[3]https://azure.microsoft.com/en-us/services/cognitive-services/bing-entity-search-api/

[4]https://www.wikidata.org/wiki/Wikidata:Main_Page

[5]https://developers.facebook.com/docs/graph-api

C. Contributions & Organization of the paper

Our contribution in this paper is two fold: (a) we propose an RDF ontology for the modelling and description of photovoltaic records with a weather component; and (b) we demonstrate a process for the conversion and storage of disparate photovoltaic and weather data to RDF data. The RDF, graph-based, data format is ideal in a data integration scenario where data, of multiple formats, is to be converted into a homogeneous form. A predefined schema (tables, columns, relationships, constraints) does not need to be created in this case, instead dynamic data records can be created on the fly. These data records can then simply be linked together to form an intricate knowledge graph. Our Photovoltaic and Weather Analysis (PWA) ontology adds the semantic layer to this knowledge graph by supplying the terms which describe the data (the photovoltaic and weather data), thus adding semantics and increasing the machine readability of the data. Our data conversion process outlines the steps involved to transform the photovoltaic and weather data to RDF, and store that data to allow its retrieval and analysis using the SPARQL query language.

The remainder of this paper is structured as follows: *Section II* presents related work and background in the areas of Linked Data, ontologies and techniques used for the conversion of non-RDF data to RDF - also known as semantic uplift; In *Section III* we detail our PWA ontology and the process we propose to semantically uplift data according to this ontology; *Section IV* presents possible use cases for our approach; *Section V* finishes the paper with conclusions and future work.

II. Related Work

A. Linked Data

Data that is made available on the Web in 'Linked Data' format [4], has great potential in that it can be easily integrated, published and connected through interlinks and Web (W3C) standards[6]. Widespread adoption of Linked Data is changing the way we use the Internet, and dramatically enhance decision-making across all sectors. Linked Data has multiple benefits[7], for example:

- Data will no longer be stored in static pieces in different places all over the Web. Instead, it will be connected and will provide greater context and a deeper understanding of data relationships (thanks to ontologies that define the schema and the meaning of data) [8].
- Data will be machine-readable, so that computers can do more 'thinking' on people's behalf by using logical inference and reasoning [9].
- Data will always be up-to-date and live federated queries can be performed online in real-time. Data is simply updated at the sources and Web (W3C) standards, such as SPARQL, allow for advanced SQL-like querying capabilities on the Web across different sources [10].
- It greatly reduces time and costs for discovering, preprocessing and integrating data [11].

B. Ontologies

A wide range of 'ontologies', alternatively called 'vocabularies', have been created in order to represent knowledge, entities and concepts for different domains [12]. These vocabularies define the schema of semantic KGs and describe their entities and relationships unambiguously. Ontologies are typically designed to be published online, reused and eventually extended[8].

In the domain of weather and renewable energy a few vocabularies exist and can be reused for our purpose. The most relevant one would be the AEMET ontology[9], a meteorology ontology network that extends the W3C Semantic Sensor Network Ontology (SSN)[10]. The aim of this ontology network is to represent knowledge related to measurements made by weather stations. The AEMET project itself, aims at making data sources from the Spanish Meteorological Office available as Linked Data [13]. The project collects data from approximately 250 automatic weather stations deployed across Spain and available as CSV files. These files are transformed (or "uplifted") to RDF data according to Linked Data principles and the AEMET ontology. While this ontology is mainly focused on weather terms and concepts, the aforementioned SSN ontology [14] is more generic and can be used as an extensible core ontology for representing any sensor observation. In this work, we take inspiration from these semantic models specifically designed for weather and sensor data and derive our own model tailored at integrating weather data with solar and renewable energy data.

In addition to the semantic models (or ontologies), a software architecture for continuously extracting and transforming this data is proposed in this paper. Strategies for making this data available online following W3C standards are described as well as use cases showing advanced querying capabilities and interlinking with different datasets (e.g. interlinking weather data with energy data or geographical information). We build on existing research, such as [15] [16] and [17]. In [15] the authors integrate weather information together with health records using Linked Data principles and a conversion process similar to ours. In [16] the authors propose guidelines for Linked Data generation and publication of building energy consumption. The SSN ontology is reused for their purpose, as data from various sensors needs to be captured. Instead, in [17] examples of data analysis on semantic sensor metadata modelled using the SSN ontology are described. In our work, we designed a core ontology (that can be aligned to the SSN and AEMET ontologies) which specifically targets the integration of weather data with PV energy data and that eases the process of mapping and uplifting common datasets used by researchers in the domain.

[6]https://www.w3.org/DesignIssues/LinkedData

[7]https://www.w3.org/standards/semanticweb/data.html

[8]https://lov.linkeddata.es/dataset/lov/

[9]http://aemet.linkeddata.es/ontology/

[10]https://www.w3.org/TR/vocab-ssn/

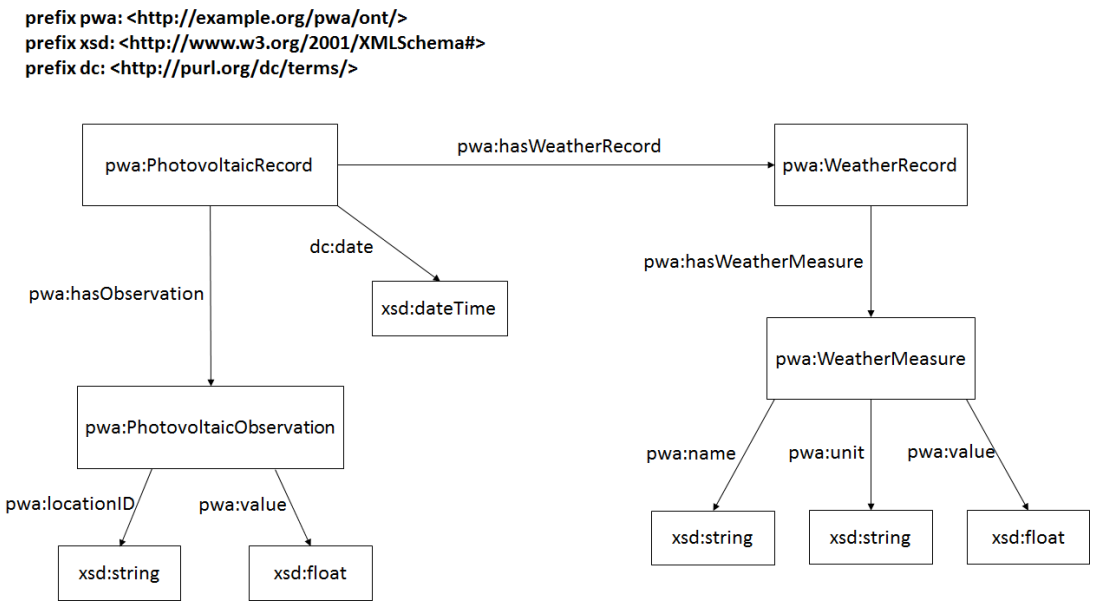

Fig. 1: The Photovoltaic and Weather Analysis (PWA) Ontology

C. Semantic Uplift

The term semantic uplift [18] is used to denote the process of converting data of a non-RDF format into RDF according to a particular ontology and is very much a classical data mapping process [19]. The purpose of such a process is to improve interoperability across data sets through transformation of data to a single format (RDF) and the addition of semantic meaning to the data - further enabling actions such as semantic search and improving the overall machine readability of the data. Semantic Uplift can be performed in many different ways, from custom scripts that perform the action as a once off task to specialised mappings languages that are used to create declarative mappings. Here we will briefly cover three approaches for semantic uplift.

The RDB to RDF Mapping Language (R2RML) [20] is a W3C recommendation for the conversion of relational and tabular data to RDF - through the creation of declarative mappings. R2RML is an expressive language that can be used to both structure the source data and describe it according to an RDF ontology. R2RML is becoming a mature standard at this stage and is already supported by some existing applications such as the Oracle and Stardog databases.

The RDF Mapping Language (RML) [21] is an extension to R2RML which supports the conversion of an increased number of data formats (Relational, CSV, TSV, JSON and XML) to RDF. RML is also a declarative mapping approach with just as much expressivity as R2RML, however, not being a W3C standard, it is less well known and it does not have the same level of support in existing applications compared to R2RML.

SPARQL-generate [22] is a declarative mapping approach to convert a multitute of data formats (Relational, CSV, TSV, JSON, XML, EXI and CBOR) into RDF. The mappings are written in a syntax that is similar to the SPARQL language, but extended with additional clauses to provide the functionality required of the mapping language. Like RML, SPARQL-generate is not a W3C standard, therefore there are less tools and applications which provide support for it.

In our data conversion process that we describe in the next section, we utilise R2RML as it is the most mature of the existing approaches.

III. PWA Ontology and Data Conversion Process

This section presents the ontology we developed to represent photovoltaic and weather data as well as the process we propose for the conversion of the original source data to RDF.

A. The Ontology

We propose a new RDF ontology, the Photovoltaic and Weather Analysis (PWA) ontology, that is specifically designed to model and describe PV data and associated weather data. The idea behind capturing both of these is that analysing PV data and weather data at a certain time and location will lead to insights into optimal operating conditions for specific PV cells and forecasting to detect these optimal conditions. The PWA ontology consists of four classes and seven properties (see Figure 1 for a visual representation of the ontology).

The *PhotovoltaicRecord* class is the main class of interest in the PWA ontology. It has a date attached to it via the *dc:date* property; it can have an arbitrary number of *PhotovoltaicObservations* attached to it by the *hasObservation* property; and a *WeatherRecord* attached via the *hasWeatherRecord* property. Note that while each photovoltaic record can have an arbitrary number of observations, each of those observations should be within close proximity of each other since there is only one weather record per photovoltaic record. It does not make sense to have two observations if those locations are 30km

apart as the weather is likely to be different in the different locations. We leave it up to the end user do determine the suitable distance between observations that will be part of one photovoltaic record.

The *PhotovoltaicObservation* class captures the energy created by a particular PV cell. Attached to each observation is the location identification number of a cell by the *locationID* property and the value of the energy generated by the cell via the *value* property.

The *WeatherRecord* class is used to model weather records which can have multiple different types of measures (such as temperature, humidity, cloud cover, wind speed, wind direction etc.). Each weather record can have an arbitrary number of *WeatherMeasures* attached to it via the *hasWeatherMeasure* property.

The *WeatherMeasure* class is used to model individual weather measures. The name of the measure (e.g. temperature) is attached via the *name* property; the unit of measurement of the measure (e.g. Fahrenheit/Celsius) is attached by the *unit* property; and the value of the measure is attached by the value property.

Listing 1: Example instance data described according to the PWA ontology.

```
@prefix pwa: <http://example.org/pwa/ont/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix dc: <http://purl.org/dc/terms/>.
@prefix data: <http://example.org/data/>.

data:1 a pwa:PhotovoltaicRecord;
  dc:date '2014-02-02T00:00:00'^^xsd:dateTime;
  pwa:hasWeatherRecord
        data:wr_2014-02-02T00:00:00;
  pwa:hasObservation _:o1.

data:wr_2014-02-02T00:00:00 a pwa:WeatherRecord;
  pwa:hasWeatherMeasure _:w1.

_:w1 a pwa:WeatherMeasure;
  pwa:name 'cloud cover'^^xsd:string;
  pwa:unit '%'^^xsd:string;
  pwa:value '10'^^xsd:float.

_:o1 a pwa:PhotovoltaicObservation;
  pwa:locationID '385'^^xsd:string;
  pwa:value '0.02079'^^xsd:float.
```

An example of data modelled according to the PWA ontology is presented in Listing 1, encoded in RDF Turtle syntax. This example contains a photovoltaic record with simply one observation and the weather record also simply has one measure.

B. The Conversion Process

In our conversion process, displayed in Figure 2, we use R2RML as the mapping language for the conversion of relational and tabular data to RDF. We choose R2RML due to its expressivity and maturity level and also because it is a declarative mapping approach. This means there is a single mapping for each piece of data to be transformed. The idea is that by taking a declarative mapping approach, maintaining the data overtime is less troublesome if the original data sources change. Since there is a mapping for each data source, it means that if a data source has changed - you simply need to find the respective mapping and update it. This way, the mapping only needs to be changed and the source code does not need to be edited and recompiled each time a data source changes.

In our conversion process - we first analyse the source data sets to see which parts of that data *aligns* with the classes and properties of the PWA ontology. When that is complete, The mappings are created which specify these alignments and the structure that the RDF is to adhere to. The R2RML mappings are then executed by the R2RML processor which takes as input the source data and the mappings and outputs the resulting RDF data. From there, this newly created RDF data is inserted into a triple-store, which is an RDF graph database, where it can now be accessed, queried and manipulated using the SPARQL query language.

IV. Use Cases

As renewable energy is becoming more and more relied upon, energy providers will further have to balance the use of renewable energy sources against traditional energy generation (coal, oil, natural gas etc.) in order to meet the demand of the electricity grid. It is envisaged that the analyses of PV data and weather data will lead to more insight regarding the electricity being generated by these cells. Through these insights, more informed decisions could be made by energy providers as to whether how heavily they can rely on the renewable energy source and cut back on the traditional energy generation - saving natural resources and reducing the production of CO2 gas. With our PWA ontology, both PV and weather data are modelled and described in a homogeneous way, facilitating greater analysis of the overall data.

Modelling instance data according to the PWA ontology will allow, through querying, some interesting retrieval and analysis use cases. Consider, for instance, the following request to an information system: *Retrieve the weather measures of all photovoltaic records where the average energy generated from the record's observations is greater than that of other photovoltaic records which have a lower cloud cover measure.* Having a system which is capable of providing an answer to such queries on top of distributed datasets may provide insight into the weather measures, other than cloud cover, which contribute to energy generation by the cells.

Another interesting retrieval use case would be, for example: *Retrieve all photovoltaic records, their energy generated and cloud cover from associated weather records where the energy generated is greater than X and the cloud cover is greater than Y* (where *X* and *Y* are two numerical values). This request has been identified by domain experts in our team as potentially relevant for the data under consideration. It serves as another example of a use case for which our proposed solution would be beneficial. We display the SPARQL query to perform this use case in Listing 2. These are just two examples of data retrieval use cases than could be posed over the data. To note that these queries could be run against multiple federated data sources in real-time and would automatically aggregate the

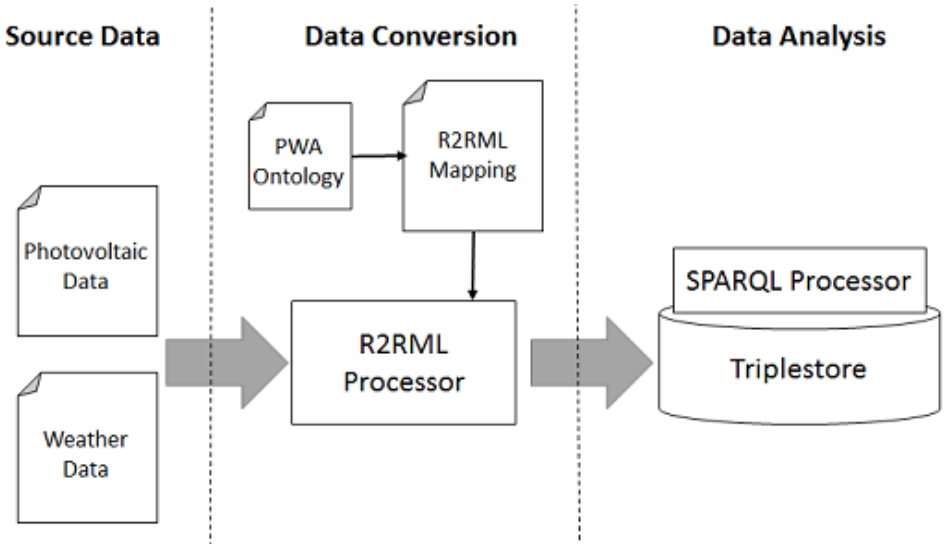

Fig. 2: Data Conversion Process

results. This is achievable simply using Linked Data standards and the proposed data modelling solution.

Listing 2: SPARQL query retrieving photovoltaic records with specific "cloud cover" and PV values (0.05 and 0.02 respectively)

```
01 PREFIX pwa: <http://example.org/pwa/ont/>
02 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
03
04 SELECT ?PVRecord ?CloudCoverValue ?PVValue
05 WHERE
06 {
07   ?PVRecord a pwa:PhotovoltaicRecord ;
08         pwa:hasWeatherRecord ?WeatherRec ;
09         pwa:hasObservation ?PVObservation .
10
11   ?WeatherRec pwa:hasWeatherMeasure ?WMeasure .
12
13   ?WMeasure pwa:name "cloud cover" ;
14         pwa:value ?CloudCoverValue .
15   FILTER (?CloudCoverValue > 0.05) .
16
17   ?PVObservation a pwa:PhotovoltaicObservation ;
18         pwa:value ?PVValue .
19   FILTER (?PVValue > 0.02) .
20 }
21 ORDER BY ?CloudCoverValue
```

V. Conclusion and Future Work

In this paper we proposed an approach for representing and converting photovoltaic and weather records into Linked Data. The goal is to make such data interoperable and homogeneous so that it can be easily harmonised and analysed across heterogeneous data sources. This would allow researchers to publish data in an interoperable format and to execute their analysis on different datasets. We proposed an RDF ontology, called PWA, for modelling and describing PV records with a weather component. Moreover, we described a process (based on R2RML) for the conversion, storage and querying of such data. In order to show the potential of the proposed solution, two concrete use cases have been detailed.

As part of our future work, we will continue working on our identified use cases (cf. Section IV). We plan to collect a large amount of data and publish it online following the described process and Linked Data principles. This data will be made available to researchers for online queries or as downloadable data dumps. Advanced queries and analysis algorithms will be implemented on top of this knowledge graph in order to test scalability, performance and compliance to requirements. Alignment with different existing ontologies and extension of the model to additional types of data will also be investigated.

References

[1] S. Manandhar, S. Dev, Y. H. Lee, and Y. S. Meng, "Analyzing solar irradiance variation from GPS and cameras," in *2018 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*. IEEE, 2018, pp. 93–94.

[2] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, "A data-driven approach to detect precipitation from meteorological sensor data," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 3872–3875.

[3] S. Manandhar, S. Dev, Y. H. Lee, S. Winkler, and Y. S. Meng, "Systematic study of weather variables for rainfall detection," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 3027–3030.

[4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic services, interoperability and web applications: emerging concepts*, pp. 205–227. IGI Global, 2011.

[5] T. Berners-Lee, J. Hendler, O. Lassila, et al., "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 28–37, 2001.

[6] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web (Dagstuhl Seminar 18371)," *Dagstuhl Reports*, vol. 8, no. 9, pp. 29–111, 2019.

[7] "Vendor landscape: Graph databases," https://go.neo4j.com/rs/710-RRC-335/images/Forrester-Research-Neo4j-Graph-Databases.pdf, Accessed: 2018-12-08.

[8] N. F. Noy, "Semantic integration: a survey of ontology-based approaches," *ACM Sigmod Record*, vol. 33, no. 4, pp. 65–70, 2004.

[9] J. Domingue, D. Fensel, and J. A. Hendler, *Handbook of semantic web technologies*, Springer Science & Business Media, 2011.

[10] S. Harris, A. Seaborne, and E. Prudhommeaux, "SPARQL 1.1 query language," *W3C recommendation*, 2013.

[11] F. Manola, E. Miller, B. McBride, et al., "RDF primer," *W3C recommendation*, 2004.

[12] P-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, "Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web," *Semantic Web*, vol. 8, no. 3, pp. 437–452, 2017.

[13] G. Atemezing, O. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, D. Vila-Suero, and B. Villazón-Terrazas, "Transforming meteorological data into linked data," *Semantic Web*, vol. 4, no. 3, pp. 285–290, 2013.

[14] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, et al., "The SSN ontology of the W3C semantic sensor network incubator group," *Web semantics: science, services and agents on the World Wide Web*, vol. 17, pp. 25–32, 2012.

[15] B. P. Reddy, B. Houlding, L. Hederman, M. Canney, C. Debruyne, C. O'Brien, A. Meehan, D. O'Sullivan, and M. A. Little, "Data linkage in medical science using the resource description framework: the AVERT model," *HRB Open Research*, vol. 1, pp. 20, mar 2019.

[16] F. Radulovic, M. Poveda-Villalón, D. Vila-Suero, V. Rodríguez-Doncel, R. García-Castro, and A. Gómez-Pérez, "Guidelines for linked data generation and publication: An example in building energy consumption," *Automation in Construction*, vol. 57, pp. 178–187, 2015.

[17] J-P. Calbimonte, O. Corcho, Z. Yan, H. Jeung, and K. Aberer, "Deriving semantic sensor metadata from raw measurements," 2012.

[18] A. Crotti Jr., C. Debruyne, R. Brennan, and D. OSullivan, "An evaluation of uplift mapping languages," *International Journal of Web Information Systems*, vol. 13, no. 4, pp. 405–424, 2017.

[19] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art," *The knowledge engineering review*, vol. 18, no. 1, pp. 1–31, 2003.

[20] S. Das, S. Sundara, and R. Cyganiak, "R2RML: RDB to RDF mapping language. W3C recommendation (2012)," 2016.

[21] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle, "RML: A generic language for integrated RDF mappings of heterogeneous data.," in *LDOW*, 2014.

[22] M. Lefrançois, A. Zimmermann, and N. Bakerally, "A SPARQL extension for generating RDF from heterogeneous formats," in *European Semantic Web Conference*. Springer, 2017, pp. 35–50.