TEC: Transparent Emissions Calculation Toolkit

Milan Markovic[1(✉)], Daniel Garijo[2], Stefano Germano[3], and Iman Naja[4]

[1] Interdisciplinary Centre for Data and AI, School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, UK
milan.markovic@abdn.ac.uk
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
daniel.garijo@upm.es
[3] Department of Computer Science, University of Oxford, Oxford, UK
stefano.germano@cs.ox.ac.uk
[4] Knowledge Media Institute, The Open University, Milton Keynes, UK
iman.naja@open.ac.uk

Abstract. Greenhouse gas emissions have become a common means for determining the carbon footprint of any commercial activity, ranging from booking a trip or manufacturing a product to training a machine learning model. However, calculating the amount of emissions associated with these activities can be a difficult task, involving estimations of energy used and considerations of location and time period. In this paper, we introduce the Transparent Emissions Calculation (TEC) toolkit, an open source effort aimed at addressing this challenge. Our contributions include two ontologies (ECFO and PECO) that represent emissions conversion factors and the provenance traces of carbon emissions calculations (respectively), a public knowledge graph with thousands of conversion factors (with their corresponding YARRRML and RML mappings) and a prototype carbon emissions calculator which uses our knowledge graph to produce a transparent emissions report.

Resource permanent URL: https://w3id.org/tec-toolkit.

Keywords: Ontology · GHG Emissions · Carbon Accounting · Transparency

1 Introduction

The Net Zero agenda has gained significant traction across the world, with over 40 countries worldwide requiring organisations to periodically calculate and report their greenhouse gas (GHG) emissions [29]. Calculating them requires real-world data observations quantifying various aspects of business activities (e.g., amount of fuel consumed by a fleet of vehicles) and additional resources such as methodologies for transforming activity data into GHG estimates (also referred to as emissions scores). Reported emissions scores may differ depending on various factors including the calculation methodology and software used,

T. R. Payne et al. (Eds.): ISWC 2023, LNCS 14266, pp. 76–93, 2023.
https://doi.org/10.1007/978-3-031-47243-5_5

geopolitical location, government requirements for reporting methods, applicable emissions conversion factors (ECFs), and the type of reported GHG emissions. Emissions calculations may also include unintentional errors, such as the use of ECFs which might be out of date, from unreliable publishers, or incorrectly applied to a specific activity, thus causing erroneous results. In addition, organisations may have a vested interest in deliberately under-reporting on certain aspects of carbon footprint if they deem it could have negative impact on the company image [20].

While reporting requirements may differ from one country to another, organisations are expected to be transparent about their submitted results. Achieving such transparency may be challenging as it requires a clear history of which ECFs were used and how the emissions scores were calculated including details about the origin and accuracy of the input data. These details are typically communicated in the form of free text reports which are not suitable for automated processing. However, such transparency is necessary to support assessments evaluating the trustworthiness and meaningful comparison of emissions scores reported by organisations across different sectors over time. We argue that provenance traces of such calculations described in the form of Knowledge Graphs (KGs) potentially provide a machine-understandable solution to this challenge by making the calculations more transparent and providing the means for automated processing and analysis. This is a core motivation for our Transparent Emissions Calculation (TEC) toolkit which aims to address this issue by providing ontologies and software tools for enhancing the transparency of emissions calculations using KGs. Our contributions include:

- Two ontologies for representing carbon emissions calculations: the Emission Conversion Factor Ontology (ECFO) and the Provenance of Emission Calculations Ontology (PECO)
- Machine-actionable mappings for transforming open ECFs data to a KG described using ECFO
- A public KG comprising of open data containing ECFs published by two different sources [10,24]
- A logic-based data validation module for the information included in the KG
- A prototype software implementation of a Semantic Machine Learning Impact calculator using various components of the TEC toolkit

The remainder of the paper is structured as follows: Sect. 2 describes related work including existing software solutions for supporting emissions calculations and semantic models considering the concept of GHG emissions; Sect. 3 describes the design of the TEC toolkit; Sect. 4 introduces ECFO and PECO ontologies, their design methodology and evaluation; Sect. 5 describes the validation of ECFO through building a public KG of ECFs; Sect. 6 describes the Semantic MLI calculator built to evaluate PECO's and ECFO's utility in the context of a software application; and Sect. 7 concludes the paper with final remarks and discussion of future work.

2 Related Work

Several ontologies have been proposed to model different aspects or sources of energy emissions, applied to different domains: in the manufacturing domain, Zhu *et al.* [42] present a carbon footprint labeling ontology to support calculation and inference of the carbon footprints of products and services (unfortunately no longer available). In [41], Zhou *et al.* present an ontology for cutting tool configuration with the goal to reduce the time, cost, and carbon footprint, while in [40] Zhang *et al.* present an ontology for modelling the carbon footprint of manufacturing mechanical products. In the logistics domain, Torres [35] describes an ontology which models performance metrics for freight transportation including safety, mobility, traffic congestion, and environment sustainability. In the built environment domain, Petri *et. al.* [31] presents an ontology that supports a digital twin model aimed at performance measurement and optimisation which include energy consumption and savings. Finally, the Open Energy Ontology[1] (OEO) [4] is a large community-led open source ontology designed for use in energy systems analysis and modelling across energy domains and to date has grown to contain over 1430 classes. The scope of these ontologies goes beyond representing conversion factor metadata, focusing on specific domains. These approaches propose mechanisms to capture emission activities (e.g., burning fuel) and consider emission factors as process attributes which qualify them. However, none of these efforts represent metadata of the corresponding ECFs used to quantify emissions or the calculations they take part in, which is the focus of our work. In fact, to the best of our knowledge, no other vocabulary focuses on capturing the provenance of emission calculations e.g., by aligning it with W3C standards (such as PROV [25] for provenance, Time [7] for applicable periods or SOSA for observations [19]) to facilitate data consumption and interoperability. A first step in this direction was presented in [17], but it was based on a more generic model and did not provide specific concepts and relations for ECFs.

From a data perspective, most countries that mandate emissions reporting require organisations to use resources provided by their governments to estimate, calculate and report emissions, with some of them providing tools to support these activities. For example, the UK government publishes an updated list of ECFs yearly as open data [10], and provides a toolkit where users can manually estimate their emissions [11]. In the US, the United States Environmental Protection Agency also publishes emission conversion factors on a yearly basis [36], along with a corresponding online platform[2] where users may enter data manually via web-forms. The GHG Protocol publishes spreadsheets to assist in the reporting of emissions by country, city, or sector [18] and the Intergovernmental Panel on Climate Change (IPCC) provides an emission factor database and application [23]. While accessing these data sources is free, working with the integrated data is challenging, as it is often made available in heterogeneous spreadsheets.

[1] https://openenergy-platform.org/ontology.
[2] https://ghgreporting.epa.gov.

Companies aim to address this data integration gap with generic (e.g., IBM [22] and Oracle [30]) and domain specific emissions calcualtors (e.g., in the logistics domain [5] and in the agriculture domain [2]). However, the opaque nature of these tools prevents their results to be easily validated or explained which becomes especially problematic in case of varying results for the same activity generated by different solutions [8,34].

Recent efforts such as Climatiq [6] have also appeared to provide API access to multi-agency ECF data in a developer-friendly format. However, these databases are not free and the APIs use application-specific request parameters (e.g., names of the units, activity identifiers) which are not defined in a formal reusable vocabulary.

3 TEC Toolkit Design

We identified the key design requirements of the TEC toolkit by assessing current UK government [10] practices for reporting ECFs and through discussions with carbon accounting experts who contributed to the development of the competency questions for our ontologies. Our requirements were also influenced by a literature review of the state of the art, in particular by current practices for measuring the impact of training AI models based on the hardware, duration and location used for training [24]. Below we summarise the main aspects of the TEC toolkit design:

Purpose: To provide the means for representing, generating, sharing, querying, and validating semantic provenance descriptions of GHG emissions calculation processes including their inputs (e.g., ECFs for different emission sources) and resulting emissions scores.

Scope: The toolkit aims to represent and map the ECFs published by different authoritative data sources with open licenses, along with their usage in emissions calculations. To narrow the scope, we targeted the UK Department for Business, Energy & Industrial Strategy (BEIS) as a data source, and calculated the provenance of ML emissions as the main activity track.

Target Users: (*1*) Organisations which report on their emissions; (*2*) Researchers in the carbon accounting domain who would like to use the toolkit or expand on its components; (*3*) Software engineers who build software to support emissions calculation processes; (*4*) Auditors and compliance officers who may use the ECF KG and other TEC toolkit software to evaluate the GHG emissions reports submitted by organisations; and (*5*) Policy-makers aiming to standardise machine-understandable GHG emissions reporting and automated carbon footprint analysis across industry sectors.

Intended Uses: (*1*) Supporting transparency by providing tools to record provenance of reported emissions scores; (*2*) Supporting the automated validations of emissions calculations to reduce errors; (*3*) Integrating data resources (e.g., ECFs) required to perform and analyse the results of emissions calculations from heterogeneous sources, and support their comparison.

Competency Questions: Based on our requirements, we created competency questions for our vocabularies, which are available online[3,4] together with example SPARQL queries and test datasets.

Governance: The TEC Toolkit is open for contributions from the community. All our repositories include contribution guidelines specifying how to propose new terms to our ontologies, new mappings for existing open datasets, or suggest new open datasets to integrate. New suggestions are addressed through GitHub issues.

4 TEC Toolkit Ontologies

The TEC toolkit contains two ontologies: the Emission Conversion Factor Ontology (ECFO)[5] and Provenance of Emission Calculations Ontology (PECO)[6]. Both ontologies are open source (CC-BY 4.0 license) and are maintained in public GitHub repositories.[7]

4.1 Methodology

We followed the Linked Open Terms methodology (LOT) [32], which proposes designing ontologies in four phases. For the first phase, the ontology requirements specification, we collected competency questions as outlined in Sect. 3. We decided to separate our ontologies in two separate vocabularies (Emission Conversion Factor Ontology (ECFO) and Provenance of Emission Calculations Ontology (PECO)), in order to ease their reusability in an independent manner. Each ontology has its corresponding requirement specification document (ORSD). ORSDs are organised as a CSV with one question per line in one column and the terms in the ontology used to address it in another column (new or reused from other vocabularies).

For the second phase (ontology implementation) we used Protégé [28] with the OWL syntax, and used LOV [37] to look for existing terms. For ECFO (see Sect. 4.2), we reused concepts from W3C PROV [25], Quantities, Units, Dimensions, and Types (QUDT) [12], Simple Knowledge Organization System (SKOS) [27], and W3C OWL-Time [7] to represent different metadata associated with an emission conversion factor. For PECO (see Sect. 4.2), we reused concepts from PROV, QUDT, and Semantic Sensor Network Ontology (SOSA) [19] ontologies to describe the provenance of emissions calculations. Ontologies were evaluated against their competency questions, as outlined in Sect. 4.4.

For the third phase (ontology publication) we generated the ontology documentations using WIDOCO [13], manually improving the results to add illustrative examples. Both ontologies have been made available using permanent URLs

[3] CQs for ECFO: https://github.com/TEC-Toolkit/ECFO/tree/main/cqs.
[4] CQs for PECO: https://github.com/TEC-Toolkit/PECO/tree/main/cqs.
[5] https://w3id.org/ecfo.
[6] https://w3id.org/peco.
[7] https://github.com/TEC-Toolkit/ECFO, https://github.com/TEC-Toolkit/PECO.

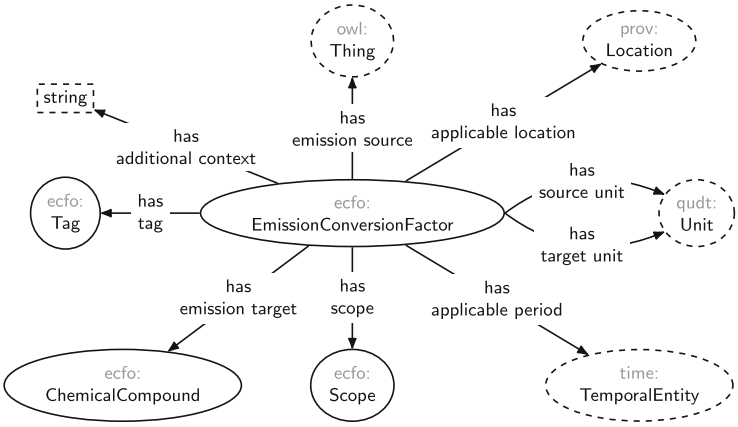

Fig. 1. Main entities of ECFO ontology. Solid ellipses represent classes, dashed ellipses refer to classes in external ontologies, dashed rectangles are RDFS datatypes and arrows represent properties.

through the w3id platform,[8] enabling content negotiation in multiple serialisations (RDF/XML, TTL, NT, JSON-LD). Each ontology has its own version IRI and a change log with the terms that differ between versions.

Finally, for the fourth phase (ontology maintenance) we tracked new issues, bugs and new requirements through GitHub issue trackers in the corresponding ontology repositories.

4.2 Emission Conversion Factors Ontology (ECFO)

ECFO aims to provide a generic model for describing the values of ECFs and their associated metadata. We represent ECFs as first class citizens, as shown in Fig. 1. Our terms have the *ecfo* prefix, while terms from imported ontologies such as W3C Time and W3C PROV use their corresponding prefixes (*time, prov*). *Ecfo:EmissionConversionFactor* represents the coefficient value used in GHG emissions calculations (i.e., activity data x emission conversion factor = GHG emissions). Each ECF instance uses *rdf:value* property to link the conversion factor value as *xsd:float*. An ECF instance is also linked (using *ecfo:hasSourceUnit*) to information about the unit of measurement (*qudt:Unit*) that corresponds to the calculation input (i.e., activity data) for which the conversion factor was designed. For example, an ECF for calculating emissions from petrol may expect the quantity of the burnt fuel to be expressed in litres. The type of the emissions source (e.g., petrol) is linked to the ECF instance via property *ecfo:hasEmissionSource*. The range of this property is *owl:Thing* as the emissions source may be conceptualised in different ways, for example, as a tangible object (e.g., fuels) but also as an event (e.g., hotel stay). An

[8] https://github.com/perma-id/w3id.org/#permanent-identifiers-for-the-web.

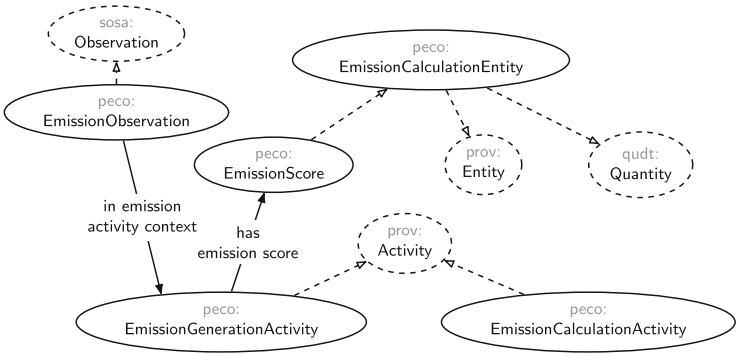

Fig. 2. Main entities of PECO ontology. Solid ellipses represent classes, dashed ellipses refer to classes in external ontologies, solid arrows represent properties and dashed arrows indicate RDFS subClassOf.

ECF may be further described using the *ecfo:Tag* concept and data property *ecfo:hasAdditionalContext* (e.g., to explain whether the ECF is considering a gross or net calorific value, or associations with specific types of machinery). Furthermore, GHG emissions are globally classified under one of three scopes [39]: direct emissions generated by activities owned or controlled by an organisation are considered Scope 1; indirect energy emissions resulting from the activities of an organisation at sources not owned or controlled by it are considered Scope 2 (e.g., purchased electricity or heat, steam and cooling) and all other indirect emissions generated by an organisation's activities at sources not owned or controlled by it are considered Scope 3 (e.g., purchased materials). ECFs may refer to the same emission source (e.g., fuel) but their values will differ significantly depending on the scope they fall under. Therefore, the term *ecfo:Scope* is used to associate each ECF instance with its relevant scope. The property *ecfo:hasTargetUnit* specifies the unit of the resulting emission factor calculated using the specific ECF (this is typically Kilograms). The kind of the emission factor (e.g., CO_2) is defined trough the *ecfo:hasEmissionTarget* property. ECFs also vary over time and the location for which they were calculated. For example, electricity generation in UK will have multiple ECFs for different years as the decarbonisation efforts of the grid progress over time. For this purpose, ECFs may use *ecfo:hasApplicablePeriod* and *ecfo:hasApplicableLocation* to further contextualise their application. Lastly, we reuse the *dc:publisher* property to link ECFs to the agent responsible for publishing the ECF values and the *prov:wasDerivedFrom* one to link an ECF to the dataset it was obtained from.

4.3 Provenance of Emission Calculation Ontology (PECO)

Figure 2 shows an overview of the PECO vocabulary, which describes provenance traces of carbon emissions calculations by capturing the quantifiable measurements of energy estimates (i.e., activity data) and ECFs used to estimate the

carbon emissions. For example, PECO helps to capture the emissions produced when electricity is consumed by machinery to manufacture a product, the petrol used to make a car journey, etc. In addition, the ontology captures data transformations that may occur before energy estimates are used with relevant ECFs.

Calculation steps are described using *peco:EmissionCalculationEntity* and *peco:EmissionCalculationActivity* which are modelled as subclasses of PROV's *prov:Entity* and *prov:Activity* respectively. A *peco:EmissionCalculationEntity* can represent any quantifiable value that was used during the calculation and the different calculation steps are linked into causal chains of events when emission calculation entities are used and generated by emission calculation activities. The *peco:EmissionCalculationEntity* is a subclass of *qudt:Quantity* to integrate mechanisms for expressing quantity values, quantity kinds and units from the QUDT ontology. The standard SOSA vocabulary is used to describe the real-world observation context (*peco:EmissionObservation*) which produces some *sosa:Result* described as *peco:EmissionCalculationEntity*. This can be observed either by a human (e.g., person reporting how long the machine was used for) or machine sensor (e.g., smart meter recording the electricity usage) and can be linked to a specific *sosa:FeatureOfInterest* (e.g., type of machine) as well as properties describing the duration of the observation period. PECO defines *peco:inEmissionActivityContext* that links the real world observations which influence emissions calculations to *peco:EmissionGenerationActivity* representing the specific activities the emissions relate to (e.g., production of goods). For convenience, the object property *peco:hasEmissionScore* may be used to link the calculated *peco:EmissionScore* to the *peco:EmissionGenerationActivity* to simplify queries that do not require to traverse the entity/activity chains documenting the full calculation process. For a more detailed example of a domain-specific emissions calculation trace annotated using PECO see Sect. 6.2.

4.4 ECFO and PECO Validation

Following our methodology, we validated both ontologies by converting their corresponding competency questions into SPARQL queries and assessing their results against real world data produced by our use cases. The CQs from ECFO have been tested against the Emission Conversion Factors KG (see Sect. 5), while the CQs from PECO have been tested against provenance traces of the Semantic Machine Learning Impact calculator (see Sect. 6). Competency questions in SPARQL and their corresponding results are available online.[9,10]

We also used the OntOlogy Pitfall Scanner (OOPS!) [33] to detect potential modeling errors, and its sister tool FOOPS! [14] to assure compliance with the Findable, Accessible, Interoperable and Reusable principles [38]. Following the feedback from these analyses, we improved our ontologies with missing metadata

[9] https://github.com/TEC-Toolkit/ECFO/blob/main/cqs/README.md.
[10] https://github.com/TEC-Toolkit/PECO/blob/main/cqs/README.md.

Table 1. Conversion factors imported in our knowledge graph, along with their applicable period and publisher.

Source	BEIS	BEIS	BEIS	BEIS	BEIS	BEIS	BEIS	MLI
Year	2022	2021	2020	2019	2018	2107	2016	2002–19
Number of ECFs	6464	6284	6140	6163	6192	6178	4977	81

and registered them in the Linked Open Vocabularies registry [37]. The reports and corresponding discussion are available online.[11,12]

5 Emission Conversion Factors Knowledge Graph

We populated ECFO with open data from two different sources. The first one is the BEIS, an open authoritative data source issuing GHG ECFs in the UK. The second source is the Machine Learning CO2 Impact Calculator [24], an open source initiative which aims to estimate the emissions of training Machine Learning models. We detail the steps followed for integrating these sources into a KG below. The resultant KG and mappings are publicly available online[13] (mappings are available under an Apache 2.0 license) [15].

5.1 Data Sources

Table 1 provides an overview of the number of ECFs in our KG, together with their publisher and publication year. In total, we include more than 42400 ECFs from diverse activities, ranging from burning fuels to driving cars of different sizes, or spending a night in a hotel.

BEIS Data: BEIS publishes annual reports of GHG emission conversion factors, together with the methodology used to estimate them. Since 2016, these data are published in a flat file CSV format, indicating the scope of the emission (Scope 1, 2, or 3), the source and target units, the type of the emission being converted to (e.g., CO_2, CH_4, etc.) as well as an up to four-level categorisation of the emission source. For example, delivery vehicles (level 1) may be vans (level 2) of certain dimensions or weight (level 3). Each combination of levels has a unique ECF value during a year, hence the high number of ECFs shown in Table 1.

In order to convert these data into a KG, we separated columns mixing units and pollutants (e.g., "kg of CO_2") and we added a column recording the valid period of time for each ECF. We then aligned units, chemical compounds and locations to Wikidata terms using Open Refine [9] and curated each result manually. Finally, we removed rows with no value for an ECF, and rows where the value was not a number. For example, we discovered that a low number of ECFs had an estimated value of "< 1". After discussing with experts, we decided not to impute a value and exclude them from our KG.

[11] ECFO reports: https://github.com/TEC-Toolkit/ECFO/issues/15.
[12] PECO reports: https://github.com/TEC-Toolkit/PECO/issues/7.
[13] https://github.com/TEC-Toolkit/cfkg.

MLI Calculator Data: The MLI calculator includes a manually curated CSV file detailing 81 different ECFs used by the tool to calculate emissions for different cloud providers around the world, separated by different compute regions (i.e., locations of data centres)[14]. We expanded this CSV file with additional information such as Wikidata IRIs corresponding to the countries associated with the compute regions, converted the ECF values for source units in kWh (i.e., to match the format in BEIS data), and where possible we also followed the referenced sources of conversion factors and extracted the applicable period range. Some source references (e.g., eGRID) were insufficiently described with missing links to the source data/tool hence we did not include these in our KG. We also fixed discrepancies between the reported conversion factors and the corresponding source references. For example, a reference used to support conversion factor value for Tokyo (Japan) only contained information about Hong Kong (China). Another reference supporting conversion factors used for Japan mentioned an ECF of 0.37kg- CO_2/kWh, however, the tool used factor 0.516. Where such discrepancies were found, we did not include applicable period range in our Knowledge Graph since the source of the ECF is unknown. We also removed the reference supporting the reported conversion factor. All reported conversion factors have been assigned a Scope 2, as they are related to electricity usage.

5.2 Transforming Data Sources to RDF

We transformed all sources using RML mappings[15] for each source file. Mappings have been developed and tested using YARRRML and Matey [21] and executed with the Morph-KGC engine [3] in RML format. All mappings include the source of the original data source in order to preserve the provenance of each ECF. For BEIS data, mapping files for each year are similar, but required small changes due to column renames in the source files. We believe that providing the mappings will help in integrating additional data sources from future years.

In total, our KG contains 662992 triples, which we expect to grow as new data sources become integrated. To help executing SPARQL queries, we have set up a public SPARQL endpoint[16] using the Fuseki Triplestore.[17]

All URIs in our KG have a permanent URL, and follow the structure:
https://w3id.org/ecfkg/i/{Region}/{Publisher}/{Year}/{cfid}

Where {*Region*}, {*Publisher*} and {*Year*} correspond to the applicable location, the responsible organisation and the year of publication of the conversion factor (respectively) and {*cfid*} corresponds to the identifier of the ECF within the source dataset.

[14] https://github.com/TEC-Toolkit/Semantic_Machine_Learning_Impact_Calculator/blob/main/src/main/resources/static/data/impact.csv.
[15] https://rml.io/specs/rml.
[16] See instructions at https://github.com/TEC-Toolkit/cfkg#sparql-endpoint.
[17] https://jena.apache.org/documentation/fuseki2.

5.3 Data Validation with Semantic Rules

We built a data validation module based on Datalog [1] rules and SPARQL queries. This allows a very high flexibility and expressivity while keeping the reasoning tractable. In addition to these advantages, Datalog has been chosen due to the wide availability of various high-performance solvers that support advanced features such as stratified negation as failure and aggregation. This way it is possible to express complex behaviours in the form of simple rules and benefit from the efficient solvers available for evaluating Datalog programs.

After loading the data and the ontology, Datalog rules are processed to infer new relations, and then `ASK` queries are used to check whether *unwanted* relations (i.e., those that represent violations of the conditions we want to validate) have any instances.[18]

By interacting with domain experts, we identified the following initial *checks* to validate:

Net/Gross CV
The *Net CV* value must be greater than or equal to the *Gross CV* value.

Kg of CO_2e
The sum of the values of all gas emissions (i.e., CO_2, CH_4 and NO_2) must be less than or equal to the *kg of* CO_2e value for the ECF referring to the same "activity".

Non-negative ECF
ECFs must be non-negative.

All these checks can be easily and naturally expressed by simple Datalog rules. For instance, the rule below identifies conversion factors (`conflictingCF`) where the *Net CV* value is less than the *Gross CV* value:

```
eov:conflictingCF(?CF_Net, ?CF_Gross) :-
    eov:sameCF(?CF_Net, ?CF_Gross) ,
    [?CF_Net, ecfo:hasAdditionalContext, "Energy - Net CV" ] ,
    [?CF_Gross, ecfo:hasAdditionalContext, "Energy - Gross CV" ] ,
    [?CF_Net, ecfo:hasEmissionTarget, ?EmissionTarget] ,
    [?CF_Gross, ecfo:hasEmissionTarget, ?EmissionTarget] ,
    ecfo:hasTargetUnit(?CF_Net, ?TargetUnit) ,
    ecfo:hasTargetUnit(?CF_Gross, ?TargetUnit) ,
    rdf:value(?CF_Net, ?Value_Net) ,
    rdf:value(?CF_Gross, ?Value_Gross) ,
    ?Value_Net < ?Value_Gross .
```

where `sameCF` is an additional relation based on properties in our ontologies to identify two ECFs that refer to the same "activity", and the prefix `eov` is used to generate unique IRIs for the data validation module.

All the details and the full set of rules used can be found in the online repository on GitHub[19] [16]. Note that the proposed approach allows to add new checks

[18] The entire process for hundreds of thousands of triples is completed in seconds on a standard laptop and uses only a few MB of RAM.

[19] https://github.com/TEC-Toolkit/Data-Validation.

just in a couple of steps[20] and effortlessly include new data to validate (all that is required is to import the relevant RDF files, and they will be automatically included in all checks).

6 Semantic Machine Learning Impact Calculator

We adapted an existing MLI calculator [24] to assess PECO when describing provenance traces of emissions calculations. We modified its code to (1) make use of our public ECF KG; (2) generate a provenance trace of the emissions calculation process, thus supporting the generation of a transparency report; and (3) include validation rules which assess whether the selected ECF is outdated, and if so to provide alternatives. This level of transparency is currently not available in other carbon footprint calculators. We named our extension the Semantic Machine Learning Impact (SMLI) calculator, and made it available online[21] [26] under MIT license.

6.1 Calculator Overview

The SMLI calculator[22] was built using a SpringBoot framework[23] with HTML and JavaScript client interface. The client extends the functionality of the original MLI calculator with the ability to document the emissions calculation trace in a JSON-LD object, fetch information about relevant ECFs from the remote KG, and evaluate the provenance trace using the application's REST services. Java-based backend services utilise Apache Jena library[24] to query a remote SPARQL endpoint and also to execute validation queries on the local in-memory model containing the provenance trace uploaded by the client.

The emissions calculation process employed by the calculator depends on user input including the location of the computation (e.g., Google Cloud Platform in asia-east1 region), hardware used (e.g., A100 PCIe 40/80 GB), and the number of hours it was used for. The subsequent emissions calculation contains two steps:

- Estimate electricity use in kWh by multiplying the hardware consumption (based on its associated thermal design power specification) and the duration of the ML model training.
- Calculate the final emission score by multiplying the estimated energy use in kWh by the relevant ECF that is applicable to the region where the ML model training took place.

Note that SMLI is suitable for some use cases more than others. While it is useful for calculating general estimates of carbon footprint for ML training, the embedded assumptions (e.g., use of thermal design power property of GPU as a proxy measurement) should be assessed by the end users before the calculator is applied in specific use cases.

[20] https://github.com/TEC-Toolkit/Data-Validation#add-another-validation-check.
[21] https://github.com/TEC-Toolkit/Semantic_Machine_Learning_Impact_Calculator.
[22] Demo: https://calculator.linkeddata.es.
[23] https://spring.io.
[24] https://jena.apache.org.

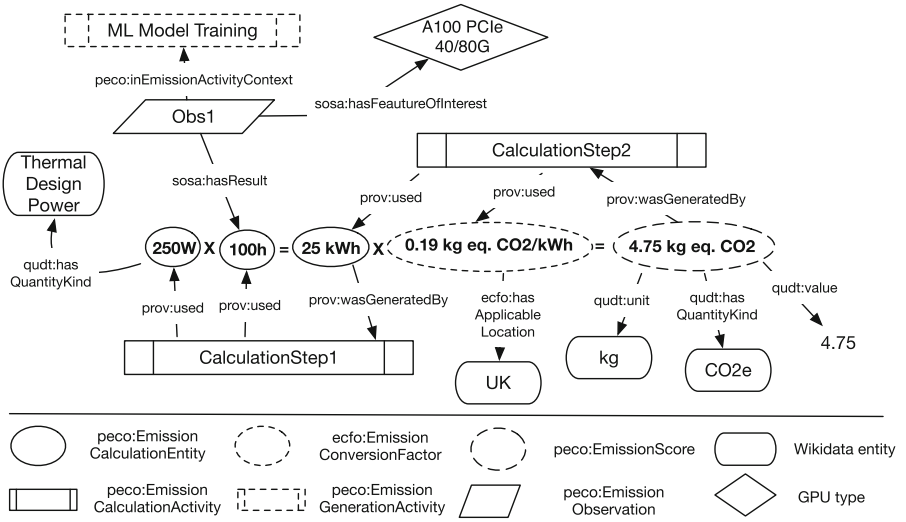

Fig. 3. An overview of the estimation (4.75 kg of CO_2e) produced by SMLI calculator (center) and its corresponding machine-readable trace aligned to PECO and ECFO. Some quantities have been simplified for clarity.

6.2 Emissions Calculation Provenance Trace

Figure 3 illustrates a portion of the provenance trace generated by the SMLI calculator. The centre of the figure depicts the aforementioned two-step emissions calculation process for training a ML model on a specific GPU hardware for 100 h. The emissions calculation process begins the with user input detailing the duration of the ML model training (*Obs1*) that links the observed value (*peco:EmissionCalculationEntity*) and the observed hardware (*sosa:FeatureOfInterest*) to the specific ML model training activity (*peco:EmissionGenerationActivity*). This is then multiplied by the thermal design power property corresponding to the specific GPU (*peco:EmissionCalculationEntity*) to produce an estimation of electricity usage in kWh which is captured as an output of the first calculation step (*peco:EmissionCalculationActivity*). The second calculation step multiplies this value with the value of the relevant conversion factor (*ecfo:EmissionConversionFactor*) to produce an estimate of the emissions released (*peco:EmissionScore*) during the ML model training. The generated provenance trace can be downloaded in a JSON-LD format.

Activity	Input	Output
Emission Score Calculation	Emission Conversion Factor - 0.19338kg [carbon dioxide equivalent]	Emission Score - 4.75E0kg [carbon dioxide equivalent]
	Energy Used - 25kWh [electricity]	
Estimate Electricity Use in kW/h	Duration of Use - 100h [time]	Energy Used - 25kWh [electricity]
	Watt Consumption - 250W [thermal design power]	

Fig. 4. Part of the transparency report detailing inputs and outputs of emission calculation activities generated from the provenance trace.

Source Unit	Target Unit	Applicable Period Start	Applicable Period End	Applicable Location	CF Value	Source	Emission Score
kWh	kg	2022-01-01T00:00:00	2022-12-31T23:59:59	United Kingdom	0.19338	link	4.75kg [carbon dioxide equivalent]
kWh	kg	2021-01-01T00:00:00	2021-12-31T23:59:59	United Kingdom	0.21233	link	5.25kg [carbon dioxide equivalent]

Fig. 5. Details of the Scope 2 electricity ECFs contained in the ECF KG and linked to the selected region of compute ordered by the applicable period. The highlighted ECF corresponds to the most recent year and is used to calculate the final emission score and included in the provenance trace.

6.3 Explaining Emissions Calculation Provenance Traces

To enhance the transparency of the calculation process, the SMLI calculator queries the provenance trace to retrieve the individual steps, inputs and outputs involved in the process (Fig. 4) including the intermediate step of estimating energy consumption based on the thermal design power property of the hardware used to perform the ML training). For each input/output the calculator shows their recorded labels, corresponding quantities with units, and the type (i.e., *qudt:QuantityKind*) of the quantity they represent.

Additionally, users are presented with a detailed overview of the ECF used to perform the calculations. The calculator is designed to query for ECF where the linked *ecfo:hasEmissionSource* value corresponds to electricity usage reported under Scope 2. For example, the information illustrated in Fig. 5 describes an ECF that converts energy expressed in kWh into kg of CO_2e and that its value is applicable to year 2022. The resulting emission score in this particular case (4.75 kg of CO_2e) is calculated based on the user input reporting a usage of

hardware with 250 W thermal design power for the period of 100 h (see Fig. 4). Where ECFs for multiple years exist in the ECF KG that correspond to the location of the compute (e.g., UK-based Google Cloud Platform europe-west2), multiple rows with corresponding information (including alternative emission scores for other years) are shown.

6.4 Assessing Provenance Traces

Semantic descriptions of emissions calculations may be further processed by validation services that help users to detect discrepancies in the calculation process, assess the quality of results, etc. To demonstrate this benefit, the software also executes SPARQL queries aimed at testing whether the ECFs used to calculate the emission score are up to date and reference the source from which they were derived. For example, to retrieve any outdated ECF recorded in the provenance trace the following SPARQL query is executed:

```
SELECT DISTINCT ?cf ?cf_value ?time
WHERE {
    ?entity a peco:EmissionScore;
        prov:wasGeneratedBy/prov:used ?cf.
    ?cf ecfo:hasApplicablePeriod/time:hasEnd/time:inXSDDate ?time;
        rdf:value ?cf_value.
    FILTER (?time < now())
}
```

If an outdated CF is selected, a warning will appear at the bottom of the page to alert users.

7 Conclusions and Future Work

In this paper we have presented the TEC toolkit, a novel ontological approach for modelling information about GHG ECFs and the provenance of GHG emissions calculations. Our toolkit includes mappings to describe two public data sources in RDF (which we have used to generate a public KG containing thousands of ECFs) and an application for estimating transparent emissions of training machine learning models built on top of our KG.

Our approach presents an open-source alternative to commercial non-semantic platforms for ECF aggregation, and demonstrates our vision of future-generation software tools producing transparent and machine-understandable records of emissions calculations that can be easily integrated and holistically analysed trough automated means. Our framework also aids in validating data consistency values, and tracking the provenance of sources the data was derived from in case corrections are needed.

In future work, we will explore how semantic descriptions of domain specific business activities (e.g., farming operations, product manufacturing, business travel, etc.) can be automatically associated with estimates of emissions source quantities (e.g., amounts of electricity/fuel used) to enable automated emission

scores calculations. Our planned use cases for the resource will initially focus on the AgriFood domain (helping UK farmers specify their farm carbon emissions accurately) and attaching provenance metadata to the emission estimates of machine learning models in Spanish Knowledge Spaces, e.g., to compare ML models by taking their emissions into account. We will also explore aligning emissions sources with existing ontologies in the energy domain (such as the Open Energy Ontology and Wikidata) in order to help intereoperability.

In addition to integrating additional open source datasets of ECFs (EPA, IPCC), we also aim to expand on the evaluations of ECFO and PECO by consulting with expert users which may lead to further extensions of these models. Finally, we would also like to explore how search and comparison of ECFs from multiple sources could be streamlined by leveraging the semantic nature of data.

Acknowledgements. This work was supported by eBay, Samsung Research UK, Siemens AG, the EPSRC projects ConCur (EP/V050869/1), UK FIRES (EP/S019111/1) and EATS (EP/V042270/1), the EU Horizon 2020 project GATEKEEPER (No 857223) and by the Comunidad de Madrid under the Multiannual Agreement with Universidad Politécnica de Madrid (UPM) in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation) and through the call Research Grants for Young Investigators from UPM. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) [or other appropriate open licence] licence to any Author Accepted Manuscript version arising from this submission.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley, Boston (1995). http://webdam.inria.fr/Alice/
2. agrecalc: Agrecalc the farm carbon calculator. https://www.agrecalc.com/. Accessed 05 May 2023
3. Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M.S., Corcho, O.: Morph-KGC: scalable knowledge graph materialization with mapping partitions. Semant. Web (2022). https://doi.org/10.3233/SW-223135
4. Booshehri, M., et al.: Introducing the open energy ontology: enhancing data interpretation and interfacing in energy systems analysis. Energy AI **5**, 100074 (2021). https://doi.org/10.1016/j.egyai.2021.100074
5. Carbon Care: CO_2 emissions calculator. https://www.carboncare.org/en/co2-emissions-calculator.html. Accessed 05 May 2023
6. Climatiq: API reference. https://www.climatiq.io/docs. Accessed 08 May 2023
7. Cox, S., Little, C.: Time Ontology in OWL. W3C candidate recommendation draft, W3C, November 2022. https://www.w3.org/TR/2022/CRD-owl-time-20221115/
8. Čuček, L., Klemeš, J.J., Kravanja, Z.: A review of footprint analysis tools for monitoring impacts on sustainability. J. Clean. Prod. **34**, 9–20 (2012)
9. Delpeuch, A., et al.: Openrefine/openrefine: Openrefine v3.7.2, April 2023. https://doi.org/10.5281/zenodo.7803000

10. Department for Energy Security and Net Zero and Department for Business, Energy & Industrial Strategy: Government conversion factors for company reporting of greenhouse gas emissions. Online at GOV.UK, June 2022. https://www.gov.uk/government/collections/government-conversion-factors-for-company-reporting
11. Department of Environment, Food & Rural Affairs: Emissions factors toolkit. Online at GOV.UK, November 2021. https://laqm.defra.gov.uk/air-quality/air-quality-assessment/emissions-factors-toolkit/
12. FAIRsharing.org: QUDT; Quantities, Units, Dimensions and Types, May 2022. https://doi.org/10.25504/FAIRsharing.d3pqw7
13. Garijo, D.: WIDOCO: a wizard for documenting ontologies. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 94–102. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_9
14. Garijo, D., Corcho, O., Poveda-Villalón, M.: Foops!: an ontology pitfall scanner for the fair principles. In: International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. CEUR Workshop Proceedings, vol. 2980. CEUR-WS.org (2021). http://ceur-ws.org/Vol-2980/paper321.pdf
15. Garijo, D., Markovic, M.: TEC-Toolkit/cfkg: CFKG 1.0.0: first release of the ECF KG, May 2023. https://doi.org/10.5281/zenodo.7916096
16. Germano, S.: TEC-Toolkit/Data-Validation: Data Validation v1.0.0, May 2023. https://doi.org/10.5281/zenodo.7916359
17. Germano, S., Saunders, C., Horrocks, I., Lupton, R.: Use of semantic technologies to inform progress toward zero-carbon economy. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 665–681. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_39
18. Green House Gas Protocol: Calculation tools. Online at ghgprotocol.org. https://ghgprotocol.org/calculation-tools. Accessed 02 May 2023
19. Haller, A., Janowicz, K., Cox, S., Phuoc, D.L., Taylor, K., Lefrancois, M.: Semantic Sensor Network Ontology. W3C recommendation, W3C, October 2017. https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/
20. He, R., Luo, L., Shamsuddin, A., Tang, Q.: Corporate carbon accounting: a literature review of carbon accounting research from the Kyoto Protocol to the Paris Agreement. Account. Finance **62**(1), 261–298 (2022). https://doi.org/10.1111/acfi.12789
21. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative rules for linked data generation at your fingertips! In: Proceedings of the 15th ESWC: Posters and Demos (2018)
22. IBM: IBM Envizi ESG Suite. https://www.ibm.com/products/envizi. Accessed 05 May 2023
23. Intergovernmental Panel on Climate Change: Efdb emission factor database, November 2020. https://www.ipcc-nggip.iges.or.jp/EFDB/main.php. Accessed 28 Apr 2023
24. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
25. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: the PROV ontology. W3C recommendation, W3C, April 2013. http://www.w3.org/TR/2013/REC-prov-o-20130430/
26. Markovic, M., Garijo, D.: TEC-Toolkit/Semantic_Machine_Learning_Impact_Calculator: SMLI Calculator 1.0.0: Stable release, May 2023. https://doi.org/10.5281/zenodo.7916120

27. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C, August 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/
28. Musen, M.A.: The protégé project: a look back and a look forward. AI Matters **1**(4), 4–12 (2015). https://doi.org/10.1145/2757001.2757003
29. Singh, N., Longendyke, L.: A global look at mandatory greenhouse gas reporting programs. Online at wri.org. https://www.wri.org/insights/global-look-mandatory-greenhouse-gas-reporting-programs. Accessed 14 Apr 2023
30. Oracle: Automate environmental data collection. https://www.oracle.com/applications/ebusiness/products/environmental-accounting-and-reporting/. Accessed 05 May 2023
31. Petri, I., Rezgui, Y., Ghoroghi, A., Alzahrani, A.: Digital twins for performance management in the built environment. J. Ind. Inf. Integr. **33**, 100445 (2023)
32. Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., García-Castro, R.: Lot: an industrial oriented ontology engineering framework. Eng. Appl. Artif. Intell. **111**, 104755 (2022). https://doi.org/10.1016/j.engappai.2022.104755
33. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: Digital twins for performance management in the built environment. Int. J. Semant. Web Inf. Syst. (IJSWIS) **10**(2), 7–34 (2014)
34. Sukhoveeva, O.: Carbon calculators as a tool for assessing greenhouse gas emissions from livestock. Dokl. Earth Sci. **497**, 266–271 (2021). Springer
35. Torres, E.J.: Ontology-driven integration of data for freight performance measures. The University of Texas at El Paso (2016)
36. United States Environmental Protection Agency: GHG emission factors hub. Online at epa.gov, April 2023. https://www.epa.gov/climateleadership/ghg-emission-factors-hub
37. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semant. Web **8**(3), 437–452 (2017)
38. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
39. World Business Council for Sustainable Development and World Resource Institute: The greenhouse gas protocol - a corporate accounting and reporting standard, revised edition. Online at ghgprotocol.org. https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf. Accessed 07 Apr 2023
40. Zhang, Y., Yi, J., Wang, Z., He, L.: A customization-oriented carbon footprint service for mechanical products. In: IOP Conference Series: Earth and Environmental Science, vol. 291, p. 012024. IOP Publishing (2019)
41. Zhou, G., Lu, Q., Xiao, Z., Zhou, C., Yuan, S., Zhang, C.: Ontology-based cutting tool configuration considering carbon emissions. Int. J. Precis. Eng. Manuf. **18**(11), 1641–1657 (2017). https://doi.org/10.1007/s12541-017-0193-2
42. Zhu, W., Zhou, G., Yen, I.L., Hwang, S.Y.: A CFL-ontology model for carbon footprint reasoning. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), pp. 224–231 (2015). https://doi.org/10.1109/ICOSC.2015.7050810