Semantic Web 0 (0) 1
IOS Press

A Semantic Approach to Reducing GHG Emissions

Kimberly Garcia [a], Jan Grau [a], Nicolas Kesseli [a], Ioannis Katis [a], Monica Arnaudo [b], Alexander Kirsten [b], Didier Beloin-Saint-Pierre [b] and Simon Mayer [a]

[a] *Department of Computer Science, University of St.Gallen, Switzerland*
E-mails: kimberly.garcia@unisg.ch, janerik.grau@student.unisg.ch, nicolas.kesseli@student.unisg.ch, ioannis.katis@student.unisg.ch, simon.mayer@unisg.ch
[b] *Technology and Society Laboratory, EMPA Swiss Federal Laboratories for Materials Science, Switzerland*
E-mails: monica.arnaudo@empa.ch, alexander.kirsten@empa.ch, didier.beloinsaintpierre@empa.ch

Abstract. In the year 2015, 196 countries signed the *Paris Agreement*, which aims at keeping the rise in mean global temperature below 2°C above pre-industrial levels. Governments have since launched awareness campaigns and tightened regulations, motivating companies and governmental organizations to reduce their direct greenhouse gas (GHG) emissions and the indirect emissions of their value chains. To monitor and report on GHG emissions, companies follow standardized methodologies which today remain costly, time-consuming, and require extensive human expertise. In this paper, we present a Knowledge Graph (KG) that forms the semantic backbone of an interdisciplinary research project that aims to significantly reduce the time and effort that environmental accounting experts spend gathering relevant data and validating it. To facilitate data gathering, instead of proposing the creation of a new standard, we created ontologies and management tools for three of the most common GHG *data formats*—ILCD, EcoSpold01, and EcoSpold02—and we propose a bridge ontology to seamlessly query data expressed in either of these formats. To take advantage of already widely-used ontologies, increase interoperability, and integrate expert knowledge, we follow the *Simplified Agile Methodology for Ontology Development* to create the WISER ontologies, which are part of the proposed KG and have been created to permit automatic responses to requests by environmental scientists and to capture their domain knowledge. To demonstrate the effectivity of our KG-based approach, we present a tool for data gathering that has been validated by environmental accounting experts. The proposed KG aims at decreasing the effort required for GHG emissions reporting while increasing its transparency and reproducibility. It furthermore democratizes access to GHG emissions data for environmental accounting experts, companies, auditing authorities, and regulatory bodies.

Keywords:
Knowledge Graph for GHG Emissions, Ontologies for Greenhouse Gas Emissions, Knowledge Graph for GHG Reporting

1. Introduction

The identification of pathways to truly and sustainably reduce the greenhouse gas (GHG) emissions of organizations and their activities requires an environmental assessment of their complete value chains, since climate change is a global issue that cannot be solved by a displacement of the problem to other regions or to a later point in time. Many assessment standards, data sources and computational tools are provided today to carry out such environmental assessments, but the abundance of options complicates the sharing and fair comparison of evaluations made by different organizations. This is an important issue for two key reasons: First, the knowledge provided by the quantitative assessment of GHG emissions from complete value chains, i.e., the carbon footprint, is mainly used to identify from a set of functionally equivalent value chains the chain that exhibits the lowest GHG emission level. This means that carbon footprint assessments are relevant only if we can compare them fairly and consistently. The second challenge comes from the fragmented and global nature of today's value chains, which are composed of diverse stakeholders such as large multi-national companies and Small to Medium Enterprises (SMEs). These organizations have to exchange consistent information on the carbon footprint of their activities Such exchanges become very difficult and time-consuming if the contexts, standards, and formats that are used for the various carbon footprint assessments are not the same. Unfortunately, this is indeed often the case, since different countries favor different assessment standards and highly informative GHG databases can be linked to unattractive costs to SMEs. Today, only environmental experts are able to deal with these challenges, which slows down the access to and integration of trustworthy GHG knowledge that could otherwise be taken advantage of by a larger part of society.

One option that is often proposed to solve these issues is the creation of a *gold assessment standard* that is accepted and used by all organizations around the world. Important efforts are made in this direction e.g., the initiatives from the Partnership for Carbon Transparency (PACT)[1] and the Together for Sustainability (TfS) initiative[2]. Until such efforts yield a common agreement, environmental accounting experts will still be needed to deal with the diverse context of carbon footprint assessment and databases to validate the relevance of evaluations from different organizations. It is therefore – today, and for the foreseeable future – very relevant to support these experts in streamlining their work by enabling automated translation and verification of GHG data for different assessment standards and contexts. Indeed, this is currently the only way to increase the relevance and trustworthiness of GHG evaluations without substantially raising the cost of carbon footprint assessments for all stakeholders of value chains.

Our interdisciplinary team, which includes environmental accounting experts, computer scientists, and business innovation specialists, proposes to tackle this pressing issue using semantic technologies to build a KG capable of: a) making data that has been described in heterogeneous yet widely accepted and (often) standardized ways accessible through a uniforming semantic layer, and b) capturing expert knowledge and integrating well-known ontologies to enable the creation of applications that facilitate data gathering for environmental accounting experts.

In this paper, we document our first major accomplishments towards this goal. After a discussion of relevant related work in Section 2, we describe the creation of ontologies for three popular environmental *data formats* – the *International Reference Life Cycle Data System* (ILCD) [1], EcoSpold01 [2], and EcoSpold02 [3]. We further discuss a bridge ontology to homogenize access to datasets expressed in these formats. In Section 4 we describe the development of the WISER ontology, which has been created following the *Simplified Agile Methodology for Ontology Development* (SAMOD)[4] to integrate well-known ontologies and capture expert knowledge to tackle specific pain point of environmental scientists. We demonstrate the proposed KG through a Web application for data gathering that has been validated by environmental accounting experts (see Section 4.2.7).

2. Related Work

We identified several relevant contributions (including published ontologies) that address sustainability at large, while only little work has focused on GHG emissions. The research that does address GHG emissions then focuses on the creation, integration, and management of datasets, ontologies, and applications for end users, while—to the best of our knowledge—there has not been research on how semantic technologies might facilitate the tasks of environmental accounting experts at the granularity we target. Our interdisciplinary work addresses this very gap, and tackles the lack of practical knowledge management approaches for GHG emissions data that are suitable for environmental accounting experts.

We discuss this related work in more detail in the following, where we have classified others' contributions in three main categories: a) creation of RDF datasets for applications that promote sustainable behaviors (Section 2.1); b) creation of knowledge models to describe environmental data (Section 2.2); and c) frameworks for environmental data and knowledge management (Section 2.3).

---

[1] https://www.carbon-transparency.com/

[2] https://www.tfs-initiative.com/

2.1. RDF Datasets for Sustainability

Wu et al., [5] focus on the integration of databases in the energy and climate sectors. They create an ontology used to semantically describe household energy consumption data and relate it to climate data. They furthermore propose tools for converting energy consumption and climate data into RDF, and make it available as Linked Data to enable the optimization of decentralized energy distribution mechanisms. KnowUREnvironment [6] proposes an unsupervised algorithm for the automatic creation of a KG that focuses on climate change and environmental issues. To create such a KG, this work utilizes 152,595 abstracts of scientific papers. Three steps are implemented: triple extraction, syntax verification with evidence counting, and graph construction. The evaluation of the resulting KG is done by asking human annotators to manually verify triple syntax, assess triple precision, and rate triple ambiguity. In [7] a Labeled Property Graph (LPG) for unit processes (i.e, smallest process elements with quantifiable inputs and outputs) is created, by combining cumulative Life Cycle Inventory (LCI) and product system datasets to enhance data interoperability. This LPG is tailored to *ecoinvent* datasets[3] [8]—the world's leading and most reputable LCI database— and automatically extract data on almost 20,000 activities (e.g., *aluminum drilling*) across more than 2,000 elementary flows (i.e., exchanges with the natural environment) stored in the LPG.

2.2. Knowledge Models for Environmental Data

Janowicz et al., [9], propose a minimal ontology design pattern to capture the key aspects of Life Cycle Assessment (LCA) and Life Cycle Inventory (LCI). In this work, the emphasis is on those data attributes that are most relevant to LCA practitioners, and the followed methodology is based on competency questions. A similar methodology is used in [10] to create a compact ontology for spatio-temporal scopes of activities in LCA. Based on this design, subsequent work [11] proposes the creation of semantic catalogs for knowledge organizations, which concisely describe heterogeneous LCA datasets to facilitate their side-by-side comparison. Although these works propose minimal patterns for LCA, they have not yet been adopted in datasets for environmental assessments. Probably due to the difficulty that shifting an entire community towards the creation of new databases that based on a new knowledge model poses.

2.3. Frameworks for Environmental Data and Knowledge Management

Konys [12] systematically analyzed 44 sustainability assessment approaches that consider social, economic and environmental factors. The main contribution is a systematic methodology for sustainability assessments knowledge, formally captured in a publicly available ontology. Closer to our objectives is Wang et al.'s framework [13], which aims at taking advantage of KGs to support environmental accounting experts conducting LCA. Three layers are proposed: *knowledge acquisition*, *knowledge graph construction*, and *applications*. The authors show a proof of concept to demonstrate the viability of the technologies. However, details on the implementation and encountered challenges are missing. Towards preserving experts knowledge, Martin et al., [14] present a systematic literature review on knowledge management in the context of sustainability. This work refers to the United Nation's Sustainable Development Goals (SDGs) and identifies two significant research gaps: a) the lack of practical approaches for knowledge management; since out of the 45 surveyed publications only one proposes a tangible and re-usable resource in the form of an ontology [12]; and b) a complete lack of action research; a methodology that investigate societal issues in collaborative teams that include practitioners and scientists [15].

2.4. Data Models for LCA

Although, the usage of Semantic Technologies for describing GHG emissions has been limited, the environmental community has already gone through great efforts in proposing ways to describe GHG data in structured ways, specifically in the context of LCA. This is driven by the continuous presence of vast amounts of produced environmental data that is described in ad-hoc ways together with the high value of achieving a consistent integration of these data, and has concretely led to the definition of formal and defacto standard—*data formats*. These include:

---

[3] https://ecoinvent.org/

- the European Commission's *International Reference Life Cycle Data System* (ILCD) [1];
- the ecoinvent *EcoSpold01*[4] and *EcoSpold02*[5] data formats;
- the GreenDelta[6] *openLCA* schema [16]; and
- the SimaPro *SimaPro-CSV*[7] data format.

While these data formats are not interoperable, EcoSpold01, EcoSpold02, and ILCD provide open source XML schemas to be used by creators of LCA datasets, and openLCA provides JSON-LD and RDF representations of their schema, while SimaPro-CSV is proprietary. Conversion software is provided by GreenDelta (open-source) and SimaPro (proprietary), and is required so the data can be used in a specific LCA computing software. This illustrates the complexity that environmental accounting experts face during their very first steps of gathering GHG data.

3. WISER: A KG for Integrated GHG Emissions Data

In our research, we focus on ILCD, EcoSpold01, and EcoSpold02, given their wide reach as the European standard for LCA data, and the data format of the leading LCI database respectively. EcoSpold01 corresponds to the data format that ecoinvent version 1 and 2 used, as well as other important databases such as Uvek [17]. EcoSpold02 is the format that ecoinvent has used for version 3 and its subversions. In the following, we describe our approach for analizing and creating RDF ontologies for EcoSpold01, EcoSpold2 and ILCD (see Sections 3.1 and 3.2). Additionally, in Section 3.3, a bridge ontology that works on top of these ontologies is proposed as a layer for providing uniform access and querying to heterogeneously described GHG data. Even though, the development of ecoinvent is based in Europe, this database is used across the world. Thus, working on these data formats provides the interdisciplinary project with a wide geographical reach.

3.1. Analyzing ILCD, EcoSpold01 and EcoSpold02

Both ecoinvent and the European Commission provide documentations on their respective data formats. When analyzing the available XML files for each format, we found out differences in naming conventions and on the definition of concepts. Both versions of EcoSpold provide a more concise description of environmental data, while ILCD divides concepts at a lower granularity and distributes them among different XML files, creating dependencies among each other. As a first step to familiarize ourselves with the formats, we manually modeled some concepts of EcoSpold02 and ILCD. The XML schema `<complexType>` tags were defined as OWL classes, and in some cases it was preferred to define them as data properties to connect classes directly instead of using identifiers. The attributes defined in a `<complexType>` tag, denoted by either the `<element>` or `<attribute>` tag were defined as data properties or object properties (following the documentation). In case the number of expected occurrences was defined either on the XML schema or on the documentation, it was modeled as the cardinality of a property. The `<documentation>` tags were modeled as annotation properties, and the `<enumeration>` tags were treated as individuals of a specific concept. Figure 1(a) shows an example of `TActivity`, which is modeled as a class related to data and object properties. A corresponding example for ILCD is presented in Figure 1 (b); specifically for the `ProcessInformationType`, which is the concept in ILCD identified as equivalent to `TActivity` in EcoSpold02. Upon inspection and confirmation with the environmental accounting experts, we found out that the ILCD data properties `nameData` and `identifierOfDataSet` are equivalent to the EcoSpold02 data properties of `activityName` and `activityNameContextId`. Moreover, the `geography` property is also common on both formats.

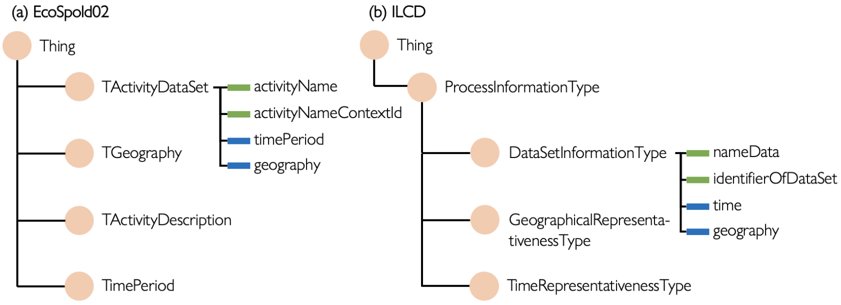


Fig. 1. A snapshot of several (corresponding) EcoSpold02 and ILCD classes created from the specification of the data formats. Classes are highlighted in light orange, object properties in blue, and data properties in green.

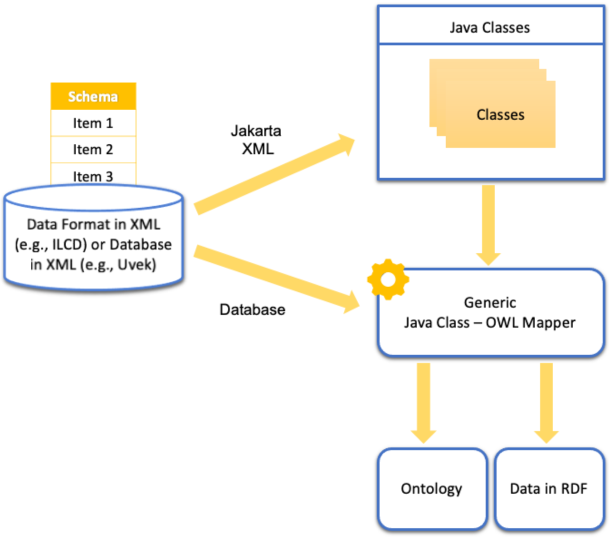


Fig. 2. Converting XML schemas into ontologies, and XML data into RDF statements.

3.2. *Automatic Creation of RDF statements from XML*

After familiarization with the considered data formats, we looked into creating tools to automatically convert XML files into RDF. Given the maturity and availability of libraries for data in XML, we decided to rely on the Jakarta XML library (formerly known as JAXB)[8]. This library allows to read the XML schema files (XSD) provided by ILCD, EcoSpold01, and Ecospold02 and convert them into their Java Class representation. The generated Java classes are a one-to-one abstraction of the given XSD file and are then used to create the RDF representation of the data format, or to create RDF statements from a database expressed in such XML data formats. Figure 2 shows an overview of this process. We followed Algorithm 1 to convert a Java class representing a dataset expressed in XML into RDF statements. We successfully implemented this approach for EcoSpold01, EcoSpold02, and ILCD data formats; and converted databases such as UVEK (EcoSpold01), Plastics Europe (ILCD), and a free version of

---

[4] https://ecoinvent.org/the-ecoinvent-database/data-formats/ecospold1/
[5] https://ecoinvent.org/the-ecoinvent-database/data-formats/ecospold2/
[6] https://greendelta.github.io/olca-schema/, GreenDelta is an independent sustainability consulting and software company.
[7] https://simapro.com/products/csv-maker/
[8] https://eclipse-ee4j.github.io/jaxb-ri/

ecoinvent (EcoSpold02) into RDF statements to be hosted in a Knowledge Graph. The transformation application is written in a generic way to support other database to be converted into RDF statements with little extra effort.

Algorithm 1 Creating RDF statements from Java Classes

```
Read a Java class $c$
Get all available properties $P$ of $c$
for $p \in P$ do
    if $p$.type() is RDF-literal then
        Create an RDF-literal
    else
        if $p$.type() is JavaClass then
            Go to 2 considering $c \leftarrow p$
        end if
    end if
end for
```

3.3. Breaking GHG Data Silos through a Bridge Ontology

To provide environmental experts with easy access to GHG data that has been expressed in different ways, instead of trying to create yet another data format—which would need to be accepted and used by the environmental community—, we propose a bridge ontology [18] that acts as an interoperability layer between data formats. The objective is to allow seamless querying of semantified data expressed in EcoSpold01, EcoSpold02, and ILCD. Such bridge ontology is based on previous analysis of environmental scientists, specifically of the openLCA project [16]. To illustrate this approach, consider the `Activity` concept in EcoSpold02 and the `ProcessDataSet` concept in ILCD. These concepts represent parts of a value chain that provide environmental accounting experts with required data for building a GHG assessment. Thus, Figure 3 shows the `TActivity` class and the `DataSetInformationType` class connected by the bridge class `BActivity`. Through `BActivity` we can hence gain uniform access to both these classes. This straightforward approach of connecting classes is not always possible, since naturally both ILCD and EcoSpold02 have different ways of describing data. However, even when classes are not equivalent, we were able to bridge objects and data properties based on the openLCA analysis [16], since some of them provide one-to-one matching, and others denote sufficiently similar properties.

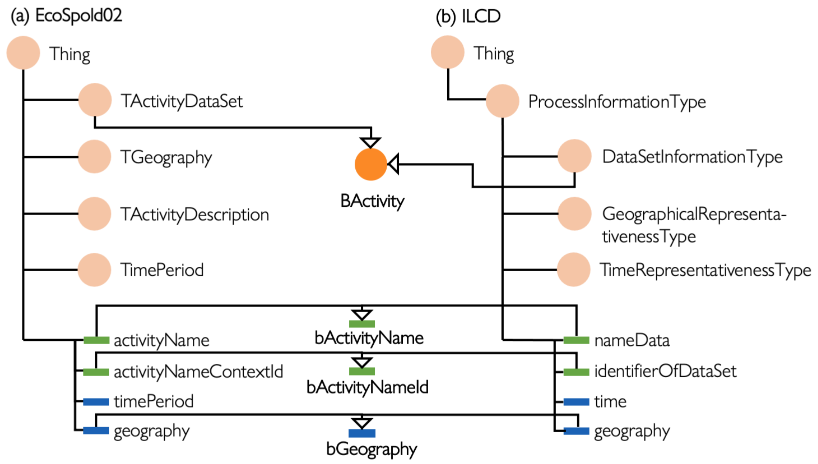


Fig. 3. Creating bridge classes between EcoSpold02 and ILCD concepts. Classes are identified in orange, object properties in blue and data properties in green.

Figure 3 hence furthermore shows a few examples of equivalent properties. The EcoSpold02 `activityName` data property is a match to the ILCD `nameData` data property, hence the `bActivityName` data property was defined. Similarly, the `bGeography` object property creates a querying bridge between the EcoSpold02 and ILCD `geography` properties, which are defined relative to different namespaces. This process was also done with EcoSpold01, notably the `TDataset` concept is aligned to the proposed `BActivity` class, and the `geography` property is matched to `bGeography` property.

4. The WISER Ontology

The bridge ontology described in the previous section allows homogeneous querying of heterogeneously described data. However, it is not capable of fulfilling all the practical requirements of environmental accounting experts when gathering data for GHG reporting. Thus, we created the WISER ontology, which integrates already available and well-known ontologies (upper as well as domain-specific) in a use-case driven manner. Moreover, domain expert knowledge is modeled in this ontology to fulfill the experts requirements. The WISER ontology was created following SAMOD [4] given that this methodology is ideally suited for collaboration between domain experts and ontology engineers, as is the case of our interdisciplinary project. We developed the WISER ontology based on a target scenario that we detail in the following.

4.1. Real-World Scenario

A large international manufacturing company has been making GHG assessments since the year 2019, revealing several challenges for carbon footprint assessment. These challenges, and the corresponding needs, have been shared with our team. One of the most relevant needs corresponds to the handling and access to valid *Scope 3* emissions data (i.e., GHG emissions emerging from companies upstream in the company's value chain) from multiple databases. Such data informs about activities in various countries and across sectors of the economy, allowing the company to make more informed decisions when modifying their value chains. The reduction of the data gathering workload and the desire for harmonization and standardization are the ultimate goals when searching for Scope 3 emissions data. Indeed, the company's experience showed that the environmental data-gathering phase is time-consuming with much uncertainty since the access to databases is often limited and it is difficult to validate the relevance of the datasets that can be found. While this experience on the one hand calls for ways to integrate data sources that has been expressed in different ways—specifically, in ILCD and the EcoSpold01, EcoSpold02 data formats—as we have discussed in the previous sections, it furthermore requires front ends that can be efficiently used by environmental accounting experts to filter and search activity datasets across heterogenous databases. Filtering is specifically relevant regarding the time and location of GHG emissions, i.e. the *location of the emissions site* and the *assessment period/year*. Ideally, a user would hence be able to select the desired filter and receive a list of compatible datasets. The ability to search across different databases and to apply filters to the available data is highly relevant to the manufacturing company, since it will allow them to reduce the amount of time and cost invested in carrying out the GHG assessments of their activities in different manufacturing sites around the world.

4.2. Building the WISER Ontology using an Agile Methodology

SAMOD [4] encourages the development of ontologies in a test-driven manner, to enable constant integration of use cases. Tackling a use case using SAMOD results in a bag of test cases T_n with sextuples consisting of: a motivating scenario (MS), a glossary of terms (GoT), a set of competency questions (CQ), a TBox-data (TBox), an ABox-data (ABox), and a set of queries (SQ). Following, we briefly describe the steps that were taken to tackle the search of datasets on a specific location as described in Section 4.1. We do not detail further the filtering of datasets by time, since it can be done using the bridge ontology, specifically the `bTimePeriod` property.

4.2.1. Motivating Scenario.

The integration of geographical interconnections can allow finding data valid for a specific context of a GHG assessment. Given that the data available in both ILCD, EcoSpold01, and EcoSpold02 only provides a string to identify the location of the data, it becomes very cumbersome for experts to search for data, not only cause they have to look for exact string matches, but also because such string matches might exclude data that can still be valid e.g., in many cases environmental accounting experts could use data that refers to a larger geographical region (*Western Europe* instead of *France*).

4.2.2. Glossary of Terms.

In order to understand domain-specific terms of the application, Table 1 shows some relevant terms for the WISER ontology. A comprehensive version of this table is available on our GitHub repository.

Term	Definition
Dataset	Corresponds to information about an activity or process that has been transformed from a *data format* such as EcoSpold01, EcoSpold02 or ILCD into RDF statements containing GHG emissions data.
Geography	A geographical region at different granularity levels, e.g. a city, a country or even a continent
GHG assessment	LCA assessment on Greenhouse Gas emission of product systems.
Activity	Term used to describe an entrepreneurial process, which has a GHG emission and is analyzed in the LCA process.
ecoInvent	Leading LCA database offering data in EcoSpold02 format.
Product System	Collection of unit processes to model the life cycle of a product. [19]
LCA	A Method to analyze environmental aspects and impacts of product systems [20].

Table 1

Glossary of Terms

4.2.3. Competency Questions.

Following we describe selected CQs that are addressed in the context of our motivating scenario (the full set of CQs is available on our GitHub repository).

CQ1. Given any region in the world: What are the available datasets in the KG that apply to this region? When searching for data of a specific region, the result should not only contain data that is valid for the exact region that a user specifies. The search results should also include data from greater regions that contain the specific region, e.g., when looking for a city—in which a specific manufacturing site is located—a search on the KG should find data that not only corresponds to the state, but also to the country, continent, and even globally.

CQ2. Given the results from CQ1, how geographically precise are the found datasets with respect to the user search?

Providing the datasets for a specific geographic region is necessary, but not sufficient. In order to create a valuable search result, the KG should be able to indicate how geographically precise the dataset is, i.e., datasets for the specific city should be better ranked than datasets for the whole continent, since the data is more exact.

4.2.4. TBox.

To address the lack of geographical interconnection, the WISER ontology has been developed relative to several best practices of the W3C Spatial Data on the Web working group [21]. For instance, re-using commonly used URI's for geographical information. Thus, we integrate the well-established GeoNames [22] ontology into our KG. Specifically, in the ABox we connected GeoNames instances with instances of the different location property values of ILCD and EcoSpold02 (e.g., *CH* or *Switzerland* link to Switzerland[9] on GeoNames). As for terminology, we

---

[9] https://www.geonames.org/2658434/

define a parent feature to connect other ontologies that contain geographical knowledge. Moreover, this WISER-specific property let us compute a geographic precision ranking. The property `bGeographyParent` is defined as:

$$\text{bGeographyParent} \sqsubseteq \text{parentFeature}^{10} \tag{1}$$

This atomic role not only allows counting the number of steps between geographical features based on the property `parentFeature` on the GeoNames ontology, but it also enables the future integration of other ontologies with a similar topological requirement.

4.2.5. ABox Part I.

The integration approach taken by the WISER ontology is based on the proposed TBox, but heavily relies on ABox assertions. In order to integrate the relevant GeoNames instances, we propose to apply a bridging instance pattern; e.g., for Switzerland:

$$\text{bGeographyTerm(BSwitzerland, "CH")} \tag{2}$$

$$\text{bGeographyTerm(BSwitzerland, "Switzerland")} \tag{3}$$

$$\text{owl:sameAs(BSwitzerland, gn:Switzerland}^{11}\text{)} \tag{4}$$

Figure 4 shows the country integration process from right to left. Once data expressed in EcoSpold02 and ILCD is converted to an RDF representation, it can be queried through our bridge ontology. At query time, the WISER ontology is used to find those instances of type `WISERGeography` whose `bGeographyTerm` property value matches the property value of the `ecoSpold:shortName` or `ilcd:location` properties. Since an instance of the `WISERGeography` class (e.g., `bSwitzerland`) is related to its equivalent instance in the GeoNames ontology, we get access to all the data that GeoNames provides. To provide a clearer understanding of our process, Figure 4 only shows the integration of GoNames with EcoSpold02 and ILCD. However, this integration has also been done for EcoSpold01 using the `geography` property.

4.2.6. ABox Part II.

Data from ecoinvent, the leading database on LCA, is expressed in EcoSpold02. However, their datasets are required to have a single geography regardless of granularity: an item may be tagged with *France* or with *European Union*, but cannot be tagged with both. This creates special cases in which datasets are assigned geographies such as “RER w/o CH & DE”, which stands for *Europe excluding Switzerland and Germany* [23]. Since these regions are not present in GeoNames, we propose the following assertions:

$$\text{bGeographyTerm(BRER\_wo\_DE\_CH, "RER wo CH \& DE")} \tag{5}$$

$$\begin{aligned}&\text{bGeographyParent}(EuropeanCountry \sqcap \forall locatedIn.Europe \sqcap \\ &\neg(BSwitzerland \sqcup BGermany) \quad \text{, BRER\_wo\_DE\_CH)}\end{aligned} \tag{6}$$

Figure 5 shows how non-standard geographical tags are handled in the WISER ontology. After data is converted to RDF, it can be queried through the bridge ontology (e.g., `BGeography`). At query time, the WISER ontology is used to find those instances of type `WISERGeography` whose `bGeographyTerm` property value matches the property value of the `ecoSpold:shortName`. To find matches for special tags, we have generated `WISERGeography` instances (e.g., `RER w/o CH & DE`) that relate through the `bGeographyParent` property to more than one instance of the type `WISERGeography` (e.g., `BSwitzerland and BFrance`), which are equivalent to a geographic region found in GeoNames.

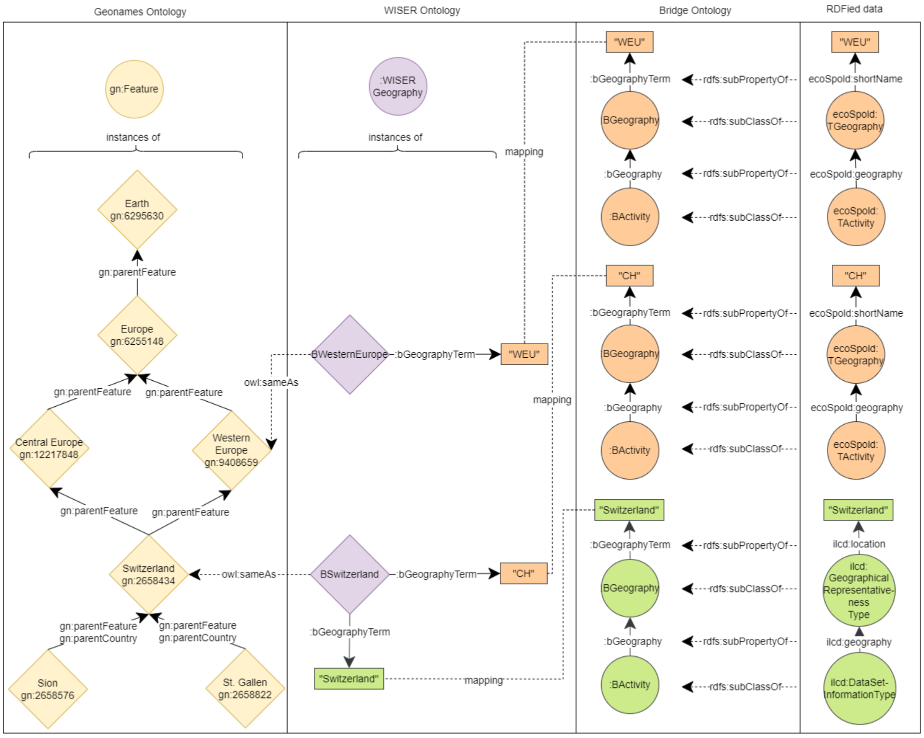


Fig. 4. Integrating GeoNames to take advantage of well-known ontologies.

4.2.7. *Set of Queries.*

The SPARQL query in Listing 1 looks for activities in the Paris region in France [12] (line 14) and an activity description containing "Electricity production" (line 15). Lines 6-9 refer to the relevant `?geographyTerm`(s) for the regions of Paris including all it's parents (using the transitive property `bGeographyParent`). The geographical precision is calculated through the `?childCounter` variable using the `bGeographyParent` property (lines 10-12). The lower the number of children steps, the more precise the result is. More queries can be found in our GitHub repository.

```
1  SELECT DISTINCT ?activity ?activityName ?geographyTerm
2  ((COUNT (DISTINCT ?childCounter)) - 1 AS ?ranking)
3  WHERE {
4      ?activity a :BActivity.
5      ?activity rdfs:label ?activityName.
6      ?activity :bGeography ?geogeography.
7      ?geography :bGeographyTerm ?geographyTerm.
8      ?filter (:bGeographyParent)* ?parent.
9      ?parent :bGeographyTerm ?geographyTerm.
10     OPTIONAL{
11         ?childCounter (:bGeographyParent)* ?parent.
12         ?filter (:bGeographyParent)* ?childCounter. }
13 #Paris region (in France)
14 FILTER(?filter = <https://sws.geonames.org/2988507/>)
```

---

[12] https://sws.geonames.org/2988507/

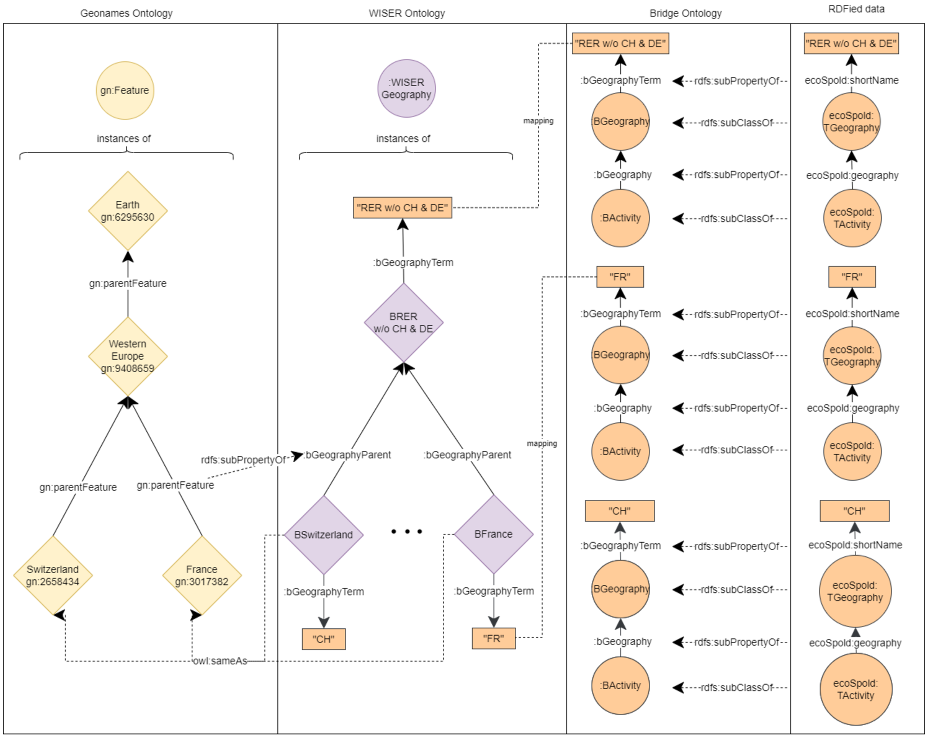


Fig. 5. Integration of geographical special cases in the WISER ontology.

```
FILTER(REGEX(?activityName, "Electricity production"))}
GROUP BY ?activity ?actvityName ?geographyTerm
ORDER BY ASC(?ranking)
```

Listing 1: Looking for datasets based on their geography

Query Performance. To evaluate the temporal cost of implementing the features tackled by the CQs, we conducted a series of benchmarking tests. Figure 6 shows the time that it took to query a KG populated with different amounts of datasets, from 10,000 to 600,000. Five queries were evaluated: the query in Listing 1 (*base query*) and four variations of it: 1) a query in which the geographies are limited to an exact match (*exact geography*), 2) a query in which the ranking for geography precision is not computed (*disable ranking*), 3) a query that receives as user input a geography that is more likely to appear in a dataset (*more likely geography*), and 4) a query that retrieves a less number of datasets given a more accurate description of the activity to look for (*reduced # of results*). The results unequivocally demonstrate that considering regions grater that a specific geography value (e.g., ask for datasets in Paris and include datasets of Western Europe) is the most costly feature.

Creating a Web applications that takes advantage of the proposed KG. Figure 7 shows a Web application created to demonstrate the proposed KG and validate that it usages can indeed make more efficient the data gathering tasks of environmental accounting experts. The highlighted dataset column ”dataset” shows that our KG allows users to query datasets that have been expressed in more than one data format; in this case from UVEK database expressed in EcoSpold01, and PlasticsEurope expressed in ILCD.

5. Conclusion and Future Work

In this paper, we report about the creation of a KG capable of acting as an interoperability layer across data that has been expressed in diverse data formats. Moreover, through the integration of well-known ontologies, the KG

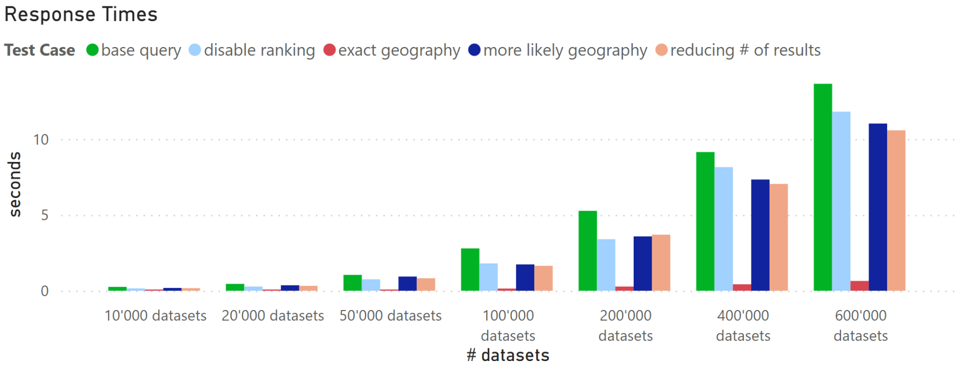
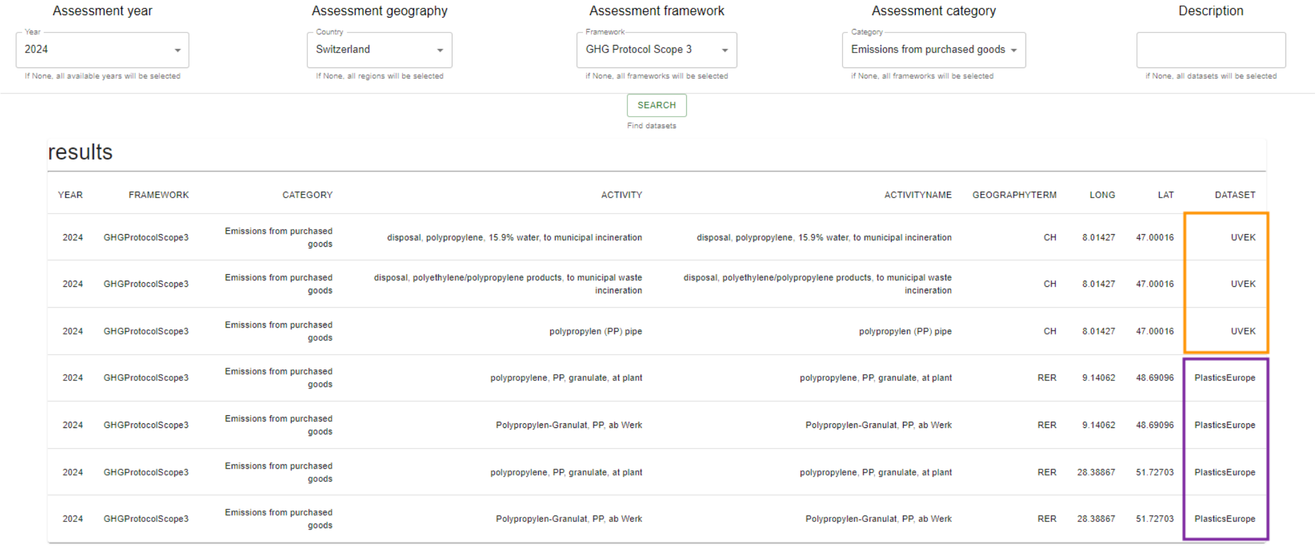


Fig. 6. Benchmark Results.



Year: 2024
If None, all available years will be selected

Country: Switzerland
If None, all regions will be selected

Framework: GHG Protocol Scope 3
if None, all frameworks will be selected

Category: Emissions from purchased goods
if None, all frameworks will be selected

if None, all datasets will be selected

results

YEAR	FRAMEWORK	CATEGORY	ACTIVITY	ACTIVITYNAME	GEOGRAPHYTERM	LONG	LAT	DATASET
2024	GHGProtocolScope3	Emissions from purchased goods	disposal, polypropylene, 15.9% water, to municipal incineration	disposal, polypropylene, 15.9% water, to municipal incineration	CH	8.01427	47.00016	UVEK
2024	GHGProtocolScope3	Emissions from purchased goods	disposal, polyethylene/polypropylene products, to municipal waste incineration	disposal, polyethylene/polypropylene products, to municipal waste incineration	CH	8.01427	47.00016	UVEK
2024	GHGProtocolScope3	Emissions from purchased goods	polypropylen (PP) pipe	polypropylen (PP) pipe	CH	8.01427	47.00016	UVEK
2024	GHGProtocolScope3	Emissions from purchased goods	polypropylene, PP, granulate, at plant	polypropylene, PP, granulate, at plant	RER	9.14062	48.69096	PlasticsEurope
2024	GHGProtocolScope3	Emissions from purchased goods	Polypropylen-Granulat, PP, ab Werk	Polypropylen-Granulat, PP, ab Werk	RER	9.14062	48.69096	PlasticsEurope
2024	GHGProtocolScope3	Emissions from purchased goods	polypropylene, PP, granulate, at plant	polypropylene, PP, granulate, at plant	RER	28.38867	51.72703	PlasticsEurope
2024	GHGProtocolScope3	Emissions from purchased goods	Polypropylen-Granulat, PP, ab Werk	Polypropylen-Granulat, PP, ab Werk	RER	28.38867	51.72703	PlasticsEurope

Fig. 7. Web application to validate the WISER ontology with environmental experts.

enables filtering and ranking capabilities of GHG data. These aspects had been identified as a large obstacle to the utilization of GHG emissions data by environmental accounting experts, especially with respect to the optimization of supply chains regarding sustainability goals. Based on three of the most common data formats for GHG emissions data–ILCD, EcoSpold01, and EcoSpold02—we created three ontologies that represent such formats, as well as a bridging ontology to permit homogeneous access to this data. Based on the concrete requirements of domain experts that were captured using competency questions (as SAMOD suggest), we created the WISER ontology and integrated GeoNames for enabling smoother geographical querying. In addition to these ontologies, which are provided as part of this publication along with the relevant SPARQL queries, we have created open-source tooling to create RDF representations of ILCD, EcoSpold01, and EcoSpold02 datasets. Finally, we have demonstrated our approach in a Web application that provides an easily usable front end for environmental experts and other users. While no formal user study has been conducted on our system, its effectivity has been verified by domain experts who especially appreciated the high experienced speed-up in finding appropriate datasets.

Furthermore, this paper provides several pathways for further work in a even broader perspective. In general two different tracks could be followed. First, we consider integrating the Shapes Constraint Language (SHACL) [24] to permit our system to verify the integrity of inserted data. Second, we consider closer collaboration with the *Partnership for Carbon Transparency* (PACT) [25] to increase the data availability for stakeholders in combination with an ontology that is able to transfer information across operational boundaries based on knowledge about the companies' relationship in the value chain (e.g., an energy supplier's *Scope 1* data could in this way automatically be discovered as relevant for a downstream company's *Scope 2* monitoring). Based on the high relevance of the challenge that our research addresses, our focus on the integration of datasets that cover a large amount of GHG emissions data, and our close and very insightful collaboration within the interdisciplinary project team and with external data providers, we believe that this research forms a valuable basis for the efficient semantics-based analysis of GHG emissions data, and thereby has a high chance to turn sustainability-oriented LCA from a manual, cumbersome, and very costly process to a seamlessly integrated feature of supply-chains world-wide.

Supplemental Material Statement: All the necessary source code to reproduce our research is available in our GitHub repository at: https://github.com/researchAndMore/swj. It includes the TTL files of the ILCD, EcoSpold02, the bridge, and wiser ontologies (including SPARQL queries), the Java code for converting XML files to RDF, the Web application demonstrator, the code for the benchmarking tests, and sample data.

References

[1] European Commission. Joint Research Centre. Institute for Environment and Sustainability., *International Reference Life Cycle Data System (ILCD) Handbook :general guide for life cycle assessment : detailed guidance.*, Publications Office, LU, 2010. https://data.europa.eu/doi/10.2788/38479.

[2] R. Frischknecht, N. Jungbluth, H.-J. Althaus, G. Doka, R. Dones, T. Heck, S. Hellweg, R. Hischier, T. Nemecek, G. Rebitzer and M. Spielmann, The ecoinvent Database: Overview and Methodological Framework (7 pp), *The International Journal of Life Cycle Assessment* **10**(1) (2005), 3–9. doi:10.1065/lca2004.10.181.1.

[3] I. Meinshausen, P. Müller-Beilschmidt and T. Viere, The EcoSpold 2 format—why a new format?, *The International Journal of Life Cycle Assessment* **21**(9) (2016), 1231–1235. doi:10.1007/s11367-014-0789-z.

[4] S. Peroni, A Simplified Agile Methodology for Ontology Development, in: *Proceedings of the 13th OWL: Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)*, 2016. https://w3id.org/people/essepuntato/papers/samod-owled2016.html.

[5] J. Wu, F. Orlandi, T. AlSkaif, D. O'Sullivan and S. Dev, A semantic web approach to uplift decentralized household energy data, *Sustainable Energy, Grids and Networks* **32** (2022), 100891. doi:https://doi.org/10.1016/j.segan.2022.100891.

[6] S. Islam, A. Proma, Y. Zhou, S.N. Akter, C. Wohn and E. Hoque, KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues, *AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges* (2022). https://www.climatechange.ai/papers/aaaifss2022/3.

[7] M. Saad, Y. Zhang, J. Tian and J. Jia, A graph database for life cycle inventory using Neo4j, *Journal of Cleaner Production* **393** (2023), 136344. doi:https://doi.org/10.1016/j.jclepro.2023.136344. https://www.sciencedirect.com/science/article/pii/S0959652623005024.

[8] G. Wernet, C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz and B. Weidema, The ecoinvent database version 3 (part I): overview and methodology, *The International Journal of Life Cycle Assessment* **21**(9) (2016), 1218–1230, Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 9 Publisher: Springer Berlin Heidelberg. doi:10.1007/s11367-016-1087-8.

[9] K. Janowicz, A.A. Krisnadhi, Y. Hu, S. Suh, P. Weidema, B. Rivela, J. Tivander, D.E. Meyer, P. Hitzler, W. Ingwersen, B. Kuczenski, Y. Ju and M. Cheatham, A Minimal Ontology Pattern for Life Cycle Assessment Data, *In: Proceedings of 6th Workshop on Ontology and Semantic Web Patterns (WOP2015) Co-located with the 14th International Semantic Web Conference (ISWC 2015).* (2015).

[10] B. Yan, Y. Hu, B. Kuczenski, K. Janowicz, A. Ballatore, A.A. Krisnadhi, Y. Ju, P. Hitzler, S. Suh and W. Ingwersen, An Ontology For Specifying Spatiotemporal Scopes in Life Cycle Assessment, in: *In: Proceedings of the Diversity Workshop Co-located with the 14th International Semantic Web Conference (ISWC 2015).*, 2015.

[11] B. Kuczenski, C.B. Davis, B. Rivela and K. Janowicz, Semantic catalogs for life cycle assessment data, *Journal of Cleaner Production* **137** (2016), 1109–1117. doi:https://doi.org/10.1016/j.jclepro.2016.07.216. https://www.sciencedirect.com/science/article/pii/S0959652616311210.

[12] A. Konys, An Ontology-Based Knowledge Modelling for a Sustainability Assessment Domain, *Sustainability* **10**(2) (2018). doi:10.3390/su10020300. https://www.mdpi.com/2071-1050/10/2/300.

[13] Y. Wang, J. Tao, W. Liu, T. Peng, R. Tang and Q. Wu, A Knowledge-enriched Framework for Life Cycle Assessment in Manufacturing, *Procedia CIRP* **105** (2022), 55–60, The 29th CIRP Conference on Life Cycle Engineering, April 4 – 6, 2022, Leuven, Belgium. doi:https://doi.org/10.1016/j.procir.2022.02.010. https://www.sciencedirect.com/science/article/pii/S2212827122000105.

[14] V.W. Martins, I. Rampasso, R. Anholon, O. Quelhas and W. Filho, Knowledge management in the context of sustainability: Literature review and opportunities for future research, *Journal of Cleaner Production* **229** (2019). doi:10.1016/j.jclepro.2019.04.354.

[15] J.M. Wittmayer, N. Schäpke, F. Van Steenbergen and I. Omann, Making sense of sustainability transitions locally: how action research contributes to addressing societal challenges, *Critical Policy Studies* **8**(4) (2014), 465–485. doi:10.1080/19460171.2014.957336.

[16] M. Srocka, J. Franze and A. Ciroth, Data conversion from ECOSPOLD02 to ILCD - openlca, GreenDeltaTC GmbH Berlin, 2010. https://www.openlca.org/wp-content/uploads/2015/11/conversion_documentation_EcoSpold02_to_ILCD_public.pdf.

[17] A. Liebich, D. Münter, S. Ludmann and T. Fröhlich, Validation of LCI data and reports of the project "Update of life cycle inventory data for natural gas and crude oil supply to Switzerland and Europe with the reference year 2019" (2019).

[18] J. Hodges, K. García and S. Ray, Semantic Development and Integration of Standards for Adoption and Interoperability, *Computer* **50**(11) (2017), 26–36. doi:10.1109/MC.2017.4041353.

[19] ISO, ISO 14040: Environmental Management - Life Cycle Assessment - Principles and Framework, ISO Central Secretariat, Geneva, 2006.

[20] W. Klöpffer and B. Grahl, *Life cycle assessment (LCA): a guide to best practice*, John Wiley & Sons, 2014.

[21] W3C Spatial Data on the Web Working Group, Spatial Data on the Web Best Practices, Technical Report, W3C Recommendation, World Wide Web Consortium, 2017, Accessed on: 2023-05-01. https://www.w3.org/TR/sdw-bp/.

[22] GeoNames, GeoNames, 2023, Accessed on: 2023-05-01. https://www.geonames.org/.

[23] ecoinvent, Geographies - ecoinvent, 2020, [Accessed May 5, 2023]. https://ecoinvent.org/the-ecoinvent-database/geographies/.

[24] World Wide Web Consortium, Shapes Constraint Language (SHACL), 2017, [Online; accessed 2-May-2023].

[25] World Business Council for Sustainable Development (WBCSD), Partnership for Carbon Transparency (PACT) sets foundations for standardized emissions data exchange, 2022, [Online; accessed 2-May-2023].

[26] The effects of climate change, NASA, 2022. https://climate.nasa.gov/effects/.

[27] Understanding Global Warming Potentials, Environmental Protection Agency, 2022. https://www.epa.gov/ghgemissions/understanding-global-warming-potentials.

[28] I.V. Muralikrishna and V. Manickam, Chapter Five - Life Cycle Assessment, in: *Environmental Management*, I.V. Muralikrishna and V. Manickam, eds, Butterworth-Heinemann, 2017, pp. 57–75. ISBN 978-0-12-811989-1. doi:https://doi.org/10.1016/B978-0-12-811989-1.00005-1. https://www.sciencedirect.com/science/article/pii/B9780128119891000051.

[29] ECOSPOLD2, Ecoinvent, 2021. https://ecoinvent.org/the-ecoinvent-database/data-formats/ecospold2/.

[30] M.A. Musen, The protégé project: a look back and a look forward, *AI Matters* **1**(4) (2015), 4–12. doi:10.1145/2757001.2757003.

[31] M.N. Asim, M. Wasim, M.U. Khan, W. Mahmood and H.M. Abbasi, A survey of Ontology Learning techniques and applications, *Database* **2018** (2018). doi:10.1093/database/bay101.

[32] I. Horrocks, Semantic Web: The Story so Far, in: *Proceedings of the 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A)*, W4A '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 120–125–. ISBN 1595935908. doi:10.1145/1243441.1243469.

[33] C. Gutierrez and J.F. Sequeda, Knowledge Graphs, *Commun. ACM* **64**(3) (2021), 96–104–. doi:10.1145/3418294.

[34] Arndt, Dörthe and Van Woensel, W., Towards supporting multiple semantics of named graphs using N3 rules, in: *13th RuleML+RR 2019 Doctoral Consortium and Rule Challenge, Proceedings*, Vol. 2438, CEUR, 2019, p. 15. ISSN 1613-0073. http://ceur-ws.org/Vol-2438/.

[35] W. Klöpffer, Life cycle assessment, *Environmental Science and Pollution Research* **4**(4) (1997), 223–228–. doi:10.1007/bf02986351.

[36] R. Frischknecht, Notions on the design and use of an ideal regional or global LCA database, *The International Journal of Life Cycle Assessment* **11**(S1) (2005), 40–48–. doi:10.1065/lca2006.04.010.

[37] W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures, W3C, 2012. https://www.w3.org/TR/xmlschema11-1/.

[38] The Sustainable Development Goals Report 2022, United Nations, N.Y., 2022. ISBN 978-92-1-101448-8. https://unstats.un.org/sdgs/report/2022/.

[39] B. Bertin, V.-M. Scuturici, E. Risler and J.-M. Pinon, A Semantic Approach to Life Cycle Assessment Applied on Energy Environmental Impact Data Management, in: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 87–94–. ISBN 9781450311434. doi:10.1145/2320765.2320796.

[40] X. Zhang, N. Heeren, C. Bauer, P. Burgherr, R. McKenna and G. Habert, The impacts of future sectoral change on the greenhouse gas emissions of construction materials for Swiss residential buildings, *Energy and Buildings* **303** (2024), 113824. doi:10.1016/j.enbuild.2023.113824. https://www.sciencedirect.com/science/article/pii/S037877882301054X.

[41] A. Mukherji, P. Thorne, W.W.L. Cheung, S.L. Connors, M. Garschagen, O. Geden, B. Hayward, N.P. Simpson, E. Totin, K. Blok, S. Eriksen, E. Fischer, G. Garner, C. Guivarch, M. Haasnoot, T. Hermans, D. Ley, J. Lewis, Z. Nicholls, L. Niamir, S. Szopa, B. Trewin, M. Howden, C. Méndez, J. Pereira, R. Pichs, S.K. Rose, Y. Saheb, R. Sánchez, C. Xiao and N. Yassaa, SYNTHESIS REPORT OF THE IPCC SIXTH ASSESSMENT REPORT (AR6) (2021). https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_SPM.pdf.

[42] T.R. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition* **5**(2) (1993), 199–220, https://pdfs.semanticscholar.org/bdaa/298f9e6eaf7d47de79c1a0e30f9436f150b6.pdf. doi:10.1006/knac.1993.1008.

[43] N.Y. Ayadi, C. Faron, F. Michel, F. Gandon and O. Corby, A Model for Meteorological Knowledge Graphs: Application to Météo-France Data, in: *22nd International Conference on Web Engineering (ICWE 2022)*, Bari, Italy, 2022. https://hal.inria.fr/hal-03619869.

[44] W3C Semantic Web Interest Group, SKOS Simple Knowledge Organization System - home page, 2023, Accessed on: 2023-05-01.

[45] Semantic Arts, GitHub - semanticarts/gist: Semantic Arts GIST Upper Enterprise Ontology, 2023, Accessed on: 2023-05-01.

[46] W3C, SPARQL Query Language for RDF, Technical Report, W3C Recommendation, World Wide Web Consortium, 2013, Accessed on: 2023-05-01. https://www.w3.org/TR/rdf-sparql-query/.

[47] U. Nations, *International Standard Industrial Classification of All Economic Activities, Revision 4*, Rev. 4th ed. edn, United Nations, New York, 2017. https://unstats.un.org/unsd/publication/seriesM/seriesM_4rev4e.pdf.

[48] World Wide Web Consortium, SKOS Simple Knowledge Organization System Reference, 2023, Accessed May 02, 2023.