DOI: 10.1111/jiec.13220

METHODS, TOOLS, DATA, AND SOFTWARE

JOURNAL OF INDUSTRIAL ECOLOGY WILEY

# A core ontology for modeling life cycle sustainability assessment on the Semantic Web

Agneta Ghose[1] | Matteo Lissandrini[2] | Emil Riis Hansen[2] | Bo Pedersen Weidema[1]

[1] Department of Planning, Danish Center of Environmental Assessment, Aalborg University, Aalborg, Denmark

[2] Department of Computer Science, Aalborg University, Aalborg, Denmark

**Correspondence**
Agneta Ghose, Danish Center of Environmental Assessment, Department of Planning, Rendsburggade 14, Aalborg University, Aalborg 9000, Denmark.
Email: agneta@plan.aau.dk

Editor Managing Review: Niko Heeren

**Funding information**
The Open Data for Sustainability Assessment project initiated by Aalborg University primarily funded this research work; Matteo Lissandrini has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement, Grant/Award Number: 838216; Emil Riis Hansen was partially funded by the Poul Due Jensen Foundation and by the Obel Family Foundation

## Abstract

The use of Semantic Web and linked data increases the possibility of data accessibility, interpretability, and interoperability. It supports cross-domain data and knowledge sharing and avoids the creation of research data silos. Widely adopted in several research domains, the use of the Semantic Web has been relatively limited with respect to sustainability assessments. A primary barrier is that the framework of the principles and technologies required to link and query data from the Semantic Web is often beyond the scope of industrial ecologists. Linking of a dataset to Semantic Web requires the development of a semantically linked core ontology in addition to the use of existing ontologies. Ontologies provide logical meaning to the data and the possibility to develop machine-readable data format. To enable and support the uptake of semantic ontologies, we present a core ontology developed specifically to capture the data relevant for life cycle sustainability assessment. We further demonstrate the utility of the ontology by using it to integrate data relevant to sustainability assessments, such as EXIOBASE and the Yale Stocks and Flow Database to the Semantic Web. These datasets can be accessed by the machine-readable endpoint using SPARQL, a semantic query language. The present work provides the foundation necessary to enhance the use of Semantic Web with respect to sustainability assessments. Finally, we provide our perspective on the challenges toward the adoption of Semantic Web technologies and technical solutions that can address these challenges.

KEYWORDS
database, industrial ecology, interoperable data, ontology, open data, Semantic Web

## 1 | INTRODUCTION

Sustainability assessment tools are data intensive and require synthesizing data from a variety of sources (Kuczenski et al., 2016). Conventional sources of data include information published in scientific publications, relevant data repositories, national/international statistics, intergovernmental organizations including UN agencies and non-governmental organizations (Cooper et al., 2013). In recent years there has been a rapid increase in the number and availability of open databases (such as EXIOBASE, USLCI, YSTAFDB) that are specifically meant to be used by researchers working with assessment tools such as life cycle and material flow assessments (Ingwersen, 2015; Merciai & Schmidt, 2018; Myers

et al., 2019b). Yet, these databases are published in different file formats (e.g., csv, xslx, xml) and also using different notations, terminologies, and schemas (Pauliuk et al., 2016), which prevents them from being interoperable immediately after publication. Therefore, researchers still face severe technical difficulties when integrating and analyzing data obtained from heterogeneous sources. Moreover, databases are periodically updated, which requires the researcher to personally follow up on how the changes in the new database version influences their models. This significantly increases the time and complexity of scientific research. These challenges make the possibility of reproducibility and reusability of projects impossible or extremely labor intensive (Ingwersen, 2015; Ingwersen et al., 2015), leading to the slowdown of scientific progress in interdisciplinary fields (Pauliuk et al., 2016). The constant increase of volume and variety of data makes it more and more necessary to use higher performance computing infrastructure beyond personal computers.

Semantic Web and Linked Open Data (LOD) are promising solutions that can be exploited to develop scalable solutions for data sharing (Fathalla et al., 2020). The Semantic Web is an extension of the current World Wide Web (WWW), in which information is given a well-defined meaning using ontologies. Ontology is a "formal description of concepts and relationships used to describe and represent data in a database"(Ragget, 2009). Ontologies or schemas enable machines and people to work in cooperation (Prud'hommeaux & Seaborne, 2008). The essential property of the WWW is its universality, that is, it does not discriminate between data sources. It is based on the power of hypertext link so that "anything can link to anything"(Berners-Lee et al., 2001). LOD represents the next step, where data on the Web is connected among different data sources (Ghali & Frayret, 2019). LOD must be in a machine-readable format, that is, data have some form of well-defined structure in contrast to natural language text, so as to be automatically searched, found, interpreted, shared, and reused across applications, organizations, and communities (Kamdar et al., 2019). An example of LOD is Dbpedia, which contains the data of the information boxes within Wikipedia (DBpedia, 2019). This data is also interlinked with several other semantically linked databases such as Geonames (geospatial information database), BiSciCol (biodiversity database), OpenEI (U.S. Department of Energy's data). Another example is Bio2RDF, a network of coherent linked data across life sciences databases such as PubMed, NCBI, Gene Ontology, and Dbpedia (Nolin et al., 2008). By linking multiple databases, the data offered becomes more complete and precise and has allowed researchers to make advances, for instance, in drug development (Kamdar et al., 2019), content management in journalism (Raimond et al., 2017), and in web search engines (Lissandrini et al., 2015; Matentzoglu et al., 2013).

Ontologies are the building blocks for describing the meaning of data in the Semantic Web (Ragget, 2009). They are particularly useful to define "knowledge" that cannot easily be represented as mathematical or statistical relationships (e.g., categorical data) (Stevens & Lord, 2009). Therefore, an ontology is a form of model that represents the key concepts of a domain and the relationships among them. Their flexibility is one of the key reasons for their widespread adoption in the field of data management for the life sciences (Yeumo et al., 2017; Kamdar et al., 2019; Stevens & Lord, 2009). The field of industrial ecology provides several tools for sustainability assessments. Among different tools, environmental assessment tools such as life cycle assessment (LCA) has gained wide acceptance and is considered essential to perform a sustainability assessment (Troullaki et al., 2021). Several attempts have been made to develop a formalized ontology particularly with respect to LCA. The CASCADE project was the first attempt to develop an ontology as a procedural guideline to develop structural classification of data collection systems in a standard LCA software (Weidema et al., 2003). However, this structuring was not in relationship to the Semantic Web. Davis et al. (2010) proposed the use of Semantic Web and LOD to collect, process, curate, and share data as a community rather than as individuals. Bertin et al. (2012) were the first to attempt the possibility of adding semantic information to life cycle inventory. In recent years other independent efforts have developed LCA ontologies in relation to semantic models (Janowicz et al., 2015; Kuczenski et al., 2016; Takhom et al., 2013; Zhang et al., 2015). Takhom (2013) and Kuczenski et al. (2016) demonstrated that structuring LCA databases using an ontology linked to the Semantic Web supports data interpretation, thus giving additional information and control to database users. Sobhkhiz et al. (2021) analyzed the challenges of managing data to perform LCA in the construction industry and highlighted how the Semantic Web is the only technology designed for dynamic data modeling. Ingwersen (2015) and Mittal et al. (2018) studied ontology development for sustainability assessment and recommended improving the linking of the semantic databases and existing ontologies to benefit from access to up-to-date information in these repositories. However, preceding studies that developed ontologies in the field of sustainability assessment did not present practical solutions with respect to the development and use of the ontologies to integrate data with the Semantic Web and LOD to enhance data interpretation and interoperability.

The conceptual framework of environmental LCA, which has traditionally focused on a limited number of ecosystem and health impacts, has been adapted to include the economic and social costs of production and consumption demand for products and processes (Guinée, 2016; Sala, 2020). This broader framework combines environmental LCA with economic life cycle costing (LCC) and social LCA (sLCA) and is referred to as life cycle sustainability assessment (LCSA). In addition, LCA is often used in combination with other tools such as MFA and input–output analysis or use complementary data to provide comprehensive analysis of impacts from production and consumption of goods and services (Crawford, Bontinck, Stephan, Wiedmann, & Yu, 2018; Lavers Westin et al., 2019; Malik, McBain, Wiedmann, Lenzen, & Murray, 2019; Pauliuk et al., 2016) . As the framework and use of sustainability assessment tools expands, the availability, variety, and the volume of data in use also increases. Existing ontologies for process-based environmental LCA do not share a common foundation with other database structures or tools used in industrial ecology (such as material flow or input–output analyses) (Pauliuk et al., 2016). Currently there exists an artificial division of data based on the different methods adopted in industrial ecology. Pauliuk et al. (2016) developed a general classification to structure data for assessing socioeconomic models commonly used in industrial ecology. They used the classification to develop a relational database, which is open source and can be used as a platform to share IE data (Pauliuk et, al., 2019). Pauliuk et al. (2016) further recommend the development of a "common collectively exhaustive ontology"

**Resource Data Framework**

Resource Data Framework is a general-purpose language to publish information on the web. It is also the fundamental specification for data representations on the Semantic Web. This framework is useful to store conceptual information about the data, that is, how one datapoint is related to another. An RDF data element is known as a triple. The structure of a triple is similar to a simple statement which contains a subject, a predicate (verb), and an object. For example: Carbon dioxide (*subject*) causes (*predicate*) Global Warming (*object*).
A subject and an object can be defined as entities, while the predicate defines the relation between the two entities. All subjects, predicates, and objects are each identified by an Internationalized Resource Identifier (IRI) which is similar to a Universal Resource Locater (URL) or a web address, hence aiding the discoverability of individual datapoints. A database of RDF triples is called a triplestore, where RDF data can be stored, and data can be retrieved using semantic queries. Semantic languages such as RDFS and OWL provide certain constructs that allow ontology developers to translate or map data into the RDF format. The primary examples of constructs used to model data into RDF format are:

- **Resource**: All things described by RDF are resources. For example, any dataset related to sustainability assessment converted into RDF would be a resource.
- **Class**: Resources are divided into groups called classes. The members of a class are known as instances of the class. For example; carbon dioxide could be an instance of a class Emission and global warming could be an instance of a class Environmental Impact.
- **Property**: A class which describes the relation between resources.
  - Domain: The classes whose instances constitute the set of resources that participate as subject in a given property.
  - Range: The classes whose instances constitute the set of resources that participate as objects in a given property.

  In the example, Carbon dioxide *causes* Global Warming; "*causes*" is a property that defines the relation between class Emission (Domain) and Environmental Impact (Range)
- **Literal**: Resources whose values are not IRIs, for example, strings, numbers, and dates.

The above components can describe hierarchies. For example, all classes are instances of Resource; all properties are instances of Class; range and domain are instances of the class Property. Classes and properties can have subclasses and subproperties. The W3C standard provides additional information on the use and structure of schemas used to represent data in RDF format (Brickley & Guha, 2014).

for the Semantic Web to support the dissemination of data and results to researchers who are unfamiliar with the terminologies used for different methods. This should make the data and results understandable and accessible to a wide range of research domains.

This study was developed within the Big Open Network for Sustainability Assessment Information (BONSAI). BONSAI is community-based organization dedicated to create a platform where all data, software, and algorithms to conduct sustainability assessments are maintained as open source (De Rosa & Weidema, 2019). The objective of this study is to enable the integration and extraction of open access data relevant to sustainability assessments with the Semantic Web. In order to do so we first develop a basic, yet comprehensive ontology that covers the core concepts in LCSA. The depth and richness of the axioms (statements and rules) depend on the aim behind the ontology development. We intend the ontology to be both specific enough to model the information contained in the data developed for LCSA and broad enough to be applicable to other data potentially useful for LCSA. Using the ontology, we give a working example of integrating and querying open access data relevant for sustainability assessments in the Semantic Web. We discuss the opportunities and challenges associated with integrating, querying, and computing data from different sources within the Semantic Web. The presented technical solutions help overcome the challenges and support the development of open data infrastructures for sustainability assessments.

## 2 | METHODOLOGY

The study involved three main phases: (i) design and development of the ontology that defines the core concepts of LCSA. (ii) Conversion of heterogeneous open datasets used for sustainability assessments into machine-readable format. (iii) Publication of the datasets on the Semantic Web with examples of querying. Each of these phases are presented in the following sections.

### 2.1 | Ontology development

The BONSAI ontology was developed, finalized, and adapted to the Resource Data Framework (RDF) (see Box 1) during two hackathons arranged in March 2019 and January 2020 with domain experts and knowledge graph developers. The ontology was developed using languages published

by the World-Wide Web Consortium (W3C) such as RDF Schema (RDFS) and Web Ontology Language (OWL) that provide basic elements for the description of ontologies, and linking as far as possible to other existing ontologies.

Based on the RDF framework, we defined classes, hierarchies, and properties for terms and attributes identified as relevant for LCSA. The principle adopted to develop the BONSAI ontology was to ensure it was general enough to be operational and usable in different domains of industrial ecology, without adhering excessively to a specific framework. To begin with, the ontology schema by Janowicz et al. (2015) was a source of inspiration in drafting the BONSAI schema. We selected terminologies for primary classes such as ***Flow***, ***Activity***, and ***Agent***. A ***Flow*** defines the transfer of an entity between different ***Activities*** performed by an ***Agent***. These three classes were selected as they define the core concepts of datasets commonly used for sustainability assessments without being too specific for a given assessment tool. The class ***Flow*** contains the measure of an entity that is produced or consumed. The entity was defined in a new class called ***FlowObject***. Unlike in the ontology developed by Janowicz et al. (2015) the ***Flow*** was not divided into economic or biosphere flows. This allows the freedom to combine economic and environmental information in a common computational structure, facilitating data exchange and reuse between different academic fields with different definitions of, for example, biosphere (Weidema et al., 2018).

We further developed new classes to fill gaps identified in the knowledge model of existing LCA ontologies and to align the terminology with the Semantic Web. For example, we introduced a class ***BalanceableProperty*** to identify quantitative flow data which follows law of conservation. We introduced a class ***DeterminingFlow*** to identify the specific flows of an activity for which a change in demand or supply will affect the activity level. A class for ***ReferenceUnit*** was introduced to annotate the common unit to which all quantitative flow data of an activity are proportional to. Note that the ***ReferenceUnit*** should not be confused with the unit corresponding to the determining flow of an activity. The ***ReferenceUnit*** of a flow is not required to be the same for all flows of the same activity but is a way to state what other measure each specific flow should be seen as proportional to, something that may depend of the original source of the flow measures. Prior to the use in calculations, disparate *ReferenceUnit*s for flow measures within the same Activity may be harmonized by transformation into measures for which the ***ReferenceUnit*** is the same amount of determining flow of an Activity. We have used the concept of class hierarchy in ontology development. For example, the class ***Flow*** contains the measure of an entity that is transferred. Hence, ***Flow*** is a subclass of the class ***Measure*** as defined by OM ontology. ***BalanceableProperty*** defines the flows that follow the law of conservation (e.g., Flows measured in wet/dry mass); ***BalanceablePropertyType*** defines the quantity of the measure (e.g., mass, energy, amount) defined in the ***BalanceableProperty*** and hence is a subclass of class ***Quantity*** as defined by OM ontology. Not all flow measurements are balanceable. Annotating non-balanceable flows is also possible since every Flow is a subclass of measure.

Furthermore, we defined the spatiotemporal information associated with ***Activity*** and ***Agent***. ***ActivityType*** defines a type of ***Activity***. An ***Activity*** is a specific occurrence of an ***ActivityType***, in a particular spatial and temporal scope. We identified and explicitly linked to other common semantic ontologies and vocabularies in order to better support data integration, discovery, and alignment. For example, for spatiotemporal properties, we link to the Time ontology in OWL (Hobbes et al., 2020) and Schema ontology (Schema Community Group, 2011; Wick & Vatant, 2012), respectively, and for units of measure we linked our ontology to the OM ontology (Rijgersberg et al., 2013). The Schema ontology was used to develop the Geonames geographical database that makes it possible to add geospatial semantic information (Wick & Vatant, 2012). Linking to this ontology enables integrating geographical data, such as names of places in various languages, geographical coordinates, and population from various sources (Wick, 2012). Geographical location includes points, lines, and spatial regions, such as cities or countries. Provenance refers to adding information on data origin. The provision of adding provenance to a dataset allows data users to verify information on when or how the data was produced. Provenance can be applied to the entire dataset as well as an individual datapoint. Adding provenance increases the transparency and trustworthiness of a dataset. We linked the BONSAI ontology to the W3C Prov-O ontology (Belhajjame et al., 2013; Hansen et al., 2020).Each of the classes are linked to other classes using one or more predicates. Twelve new predicates were also defined in the BONSAI ontology. Details on the primary classes and predicates in the BONSAI ontology are given in Tables 1 and 2. The finalized ontology defines all the concepts of LCSA in the form of triples (see Figure 1). The ontology was published in a BONSAI-specific namespace.[1] Specific IRIs were created for each class and predicate defined in the ontology. Once the ontology was finalized, we proceeded to extract and annotate data from publicly available open datasets in order to link the dataset to the Semantic Web following our ontology and then to validate it.

## 2.2 | Data extraction

The objective of Semantic Web and LOD is to provide a common framework that allows data to be shared and reused across applications. There are numerous datasets available for sustainability assessments, however the access to a large share of established datasets is proprietary. To test the ontology, we chose two established open-access datasets.

**EXIOBASE** is a multi-regional input–output (MRIO) model that includes supply and use tables for 44 countries, 5 Rest of World regions, 200 products, and 164 industries/activities in each geographical area (EXIOBASE Consortium, 2014; Merciai & Schmidt, 2018). In addition, the dataset includes accounts on environmental emissions, waste, use of land and natural resources for each activity. EXIOBASE supports analysis techniques

[1] https://ontology.bonsai.uno/

**TABLE 1** Classes introduced in the BONSAI ontology aligned to the Semantic Web Here "bont:" is the suggested prefix for the namespace for the BONSAI ontology IRI[1]

| Classes | Description | Example instances |
|---|---|---|
| bont: Activity | Making or doing something within a spatial and temporal delimitation. This is one of the identifying dimensions of a datapoint. This class defines multiple properties on the type and direction of flows. "Process" is a commonly used as a synonym in other LCA databases. | For example, "Cultivation of wheat" in Germany in the year 2020 or "Aluminium production" in China in the year 2020. |
| bont: ActivityType | This defines the type of an activity. This class includes the labels of activities. Includes both human activities (e.g., activities for production and consumption, market activities, and stock accumulation activities) and environmental mechanisms (e.g., atmospheric energy balance, deposition, pollination). | In EXIOBASE, examples of *ActivityTypes* are "Cultivation of wheat" or "Aluminum production." In YSTAFDB, examples of *ActivityTypes* are "societal consumption" or "fabrication and manufacturing." |
| bont: Flow | An input or output of an entity to or from an instance of an *Activity* or a directional exchange of an entity between two instances of *Activity*. In bipartite graph theory, a Flow would be an edge of the graph, while Activity would be vertices. However, a flow can also be unidirectional. That is, a flow can be defined as an input or output of an activity without defining its origin or destination. | An example of a Flow, in EXIOBASE would be the input of 2393 tonnes of "Aluminium and aluminum products" (*FlowObject*) to "Manufacture of motor vehicles" (*ActivityType*) in Germany in the year 2011.<br>In YSTAFDB, an output of 2684 megagram (tonnes) of "jewelry and silverware" from silver "fabrication and manufacturing" (*ActivityType*) in India in the year 1997. |
| bont: FlowObject | This class includes the labels of entities that are produced or consumed by an activity or added to or removed from a stock accumulation. | In EXIOBASE, examples of *FlowObjects* are "Wheat" or "Aluminum and aluminium products." In YSTAFDB, examples of *FlowObjects* are "phosphate rock output" or "phosphorus, societal consumption output."<br>FlowObjects also include names of natural resources (land, minerals, etc.); waste or emissions as well as non-material flows such as social or economic flows (employment, taxes, compensation, etc.) |
| bont: BalanceableProperty | A measure that follows a conservation law. In a complete description of an activity, the sum of all measures for all input flows must equal the sum of all measures for all output flows, when all these measures are expressed in the same unit. Balanceable properties are particularly relevant for validating the completeness and consistency of an Activity description or a database of such activities. | Common examples of BalanceableProperty are Mass and Energy. For example, dry mass, wet mass, energy, elemental mass, and monetary value when measured in the same valuation.<br>Examples of quantities that are not balanceable are—volume, number of units, Becquerel (unit to measure radioactivity). Instances of non-balanceable properties could be converted to balanceable properties with additional information, for example, the density of an instance of flow can be used to convert its volume to mass units. |
| bont: BalanceablePropertyType | The BalanceablePropertyType defines the quantity of the measure defined in the BalanceableProperty. | Examples of BalanceableProperty are dry mass, wet mass, energy, elemental mass. The BalanceablePropertyType would be Mass. |
| bont: Agent | An entity (person or thing) that performs an activity. An agent may have a location that may be different from the location of an Activity performed by it. | Within an activity, agents can perform different roles, for example, laborer, owner, purchaser, consumer |
| bont: ReferenceUnit | A measure to which the numeric value representing the measure of a Flow is expressed in proportion to. In LCA, the term "Functional Unit" is defined as a common ReferenceUnit for all activities in an LCA study. "Functional Units" are ReferenceUnits, but not all ReferenceUnits are "Functional Units." | For example, the amount of $CO_2$ emitted from a transport activity may be expressed in proportion to the quantity of another flow of this activity (e.g., 1 km of distance covered) or to a time period (e.g., $CO_2$ emissions per year from transport). |

[1] https://ontology.bonsai.uno/core/ontology_v0.2.ttl

such as LCA and input output analysis to analyze the environmental pressures of economic activities (Beylot et al., 2019; Schmidt & De Rosa, 2020; Schmidt et al., 2021). All versions of EXIOBASE can be downloaded from the official website (Exiobase Consortium, 2014). In this study, we used the hybrid version of EXIOBASE version 3.7.17 which contains the MRIO model from 2011. Supply and Use tables in EXIOBASE have a format of products by activities.

**TABLE 2** Predicates introduced in the BONSAI ontology aligned to the Semantic Web Here "bont:," "time:" and "schema:" refer to the suggested prefixes for the namespaces for BONSAI ontology IRI[1], OWL time ontology[1], and the Schema ontology,[2] respectively

| Predicate | Description | Domain | Range |
|---|---|---|---|
| **bont: isInputOf** | Specifies the Activity that a Flow is an input to | bont: Flow | bont: Activity |
| **bont: isOutputOf** | Specifies the Activity that a Flow is an output of | bont: Flow | bont: Activity |
| **bont: hasObjectType** | Specifies the Flow Object consumed or produced | bont: Flow | bont: *FlowObject* |
| **bont: hasActivityType** | Specifies the type of the Activity | bont: Activity | bont: ActivityType |
| **bont: hasDeterminingFlow** | Specifies a flow object produced or consumed by an activity for which a change in demand or supply will affect the activity level (such as its production volume or extent) | bont: Activity | bont: Flow |
| **bont: performs** | Specifies the Activity that an Agent performs | bont: Agent | bont: Activity |
| **bont: hasTemporalExtent** | Specifies the temporal extent of an Activity | bont: Activity | time: ProperInterval |
| **bont: hasLocation** | Specifies the location of an Activity or Agent | bont: Activity, bont: Agent | schema:location |
| **bont: isProportionalTo** | Specifies the reference unit that the amount of a BalanceableProperty of a Flow is proportional to | bont: BalanceableProperty | bont: ReferenceUnit |
| **bont: hasBalanceableProperty** | Specifies the Measure of a Flow when this Measure is a BalanceableProperty, that is, when it follows a conservation law | bont: Flow | bont: BalanceableProperty |
| **bont: hasPropertyType** | Specifies the dimension (Quantity) of a Measure that is classified as a BalanceableProperty | bont: BalanceableProperty | bont: BalanceablePropertyType |
| **bont: hasReference*FlowObject*** | Specifies a FlowObject that functions as the ReferenceUnit for a BalanceableProperty measure | bont: ReferenceUnit | bont: *FlowObject* |

[1] https://www.w3.org/2006/time#
[2] https://schema.org/

**YSTAFDB** was developed from material stocks and flows data generated at the Center for Industrial Ecology at Yale University (Myers et al., 2019b). This contains 100,000+ data records on anthropogenic cycles of 62 elements as well as specific engineering materials such as stainless steel. This dataset is published as a supplementary material on zenodo.org alongside the main publication (Myers et al., 2019a, 2019b).

Both datasets can be used for sustainability assessments individually or in combination. Most datasets are shared in non-normative formats. For example, the EXIOBASE dataset is shared as a set of excel spreadsheets and YSTAFDB datasets are provided as plain text CSV files. To allow automatic transformation and integration of datasets by a common set of data converters, we developed a common intermediate CSV format. This intermediate CSV format formalizes a list of instances of *Flows*, *FlowObjects*, *ActivityTypes*, and *Locations* in a given dataset.

The datasets were annotated using the BONSAI ontology and converted into RDF using software developed by the BONSAI community based on RDFlib.[2,3] Each datapoint gets its own IRI. For EXIOBASE, the terms ":products," "resources," and "emissions" were annotated as *FlowObjects*; whilst "economic activities" including consumption at final demand was annotated as *ActivityTypes*. EXIOBASE has locations referred to as "RoW Asia and Pacific." This is not a formalized description of a geographic location and hence unavailable in the instances given in Schema ontology. For such datapoints, specific IRI(s) were created to ensure this information can still be converted to RDF. The structure of the EXIOBASE supply table matrix is such that flows on the diagonal cells of the matrix are implicitly the *DeterminingFlow for the Activity*, while other dependent flows (often referred as by-products) are in cells outside the diagonal.

[2] The software and code to convert YSTAFDB data to RDF format are available at https://github.com/BONSAMURAIS/ystafdb
[3] The software and code to convert EXIOBASE data to RDF format are available at https://github.com/BONSAMURAIS/EXIOBASE-conversion-software

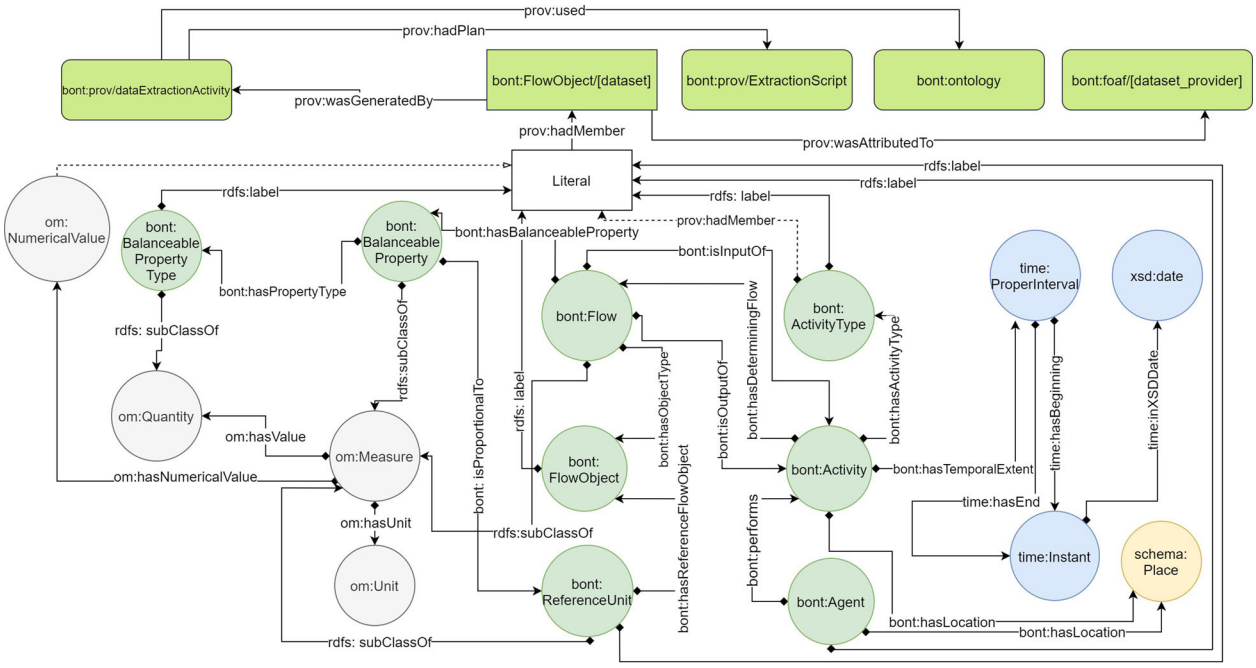


**FIGURE 1** Diagrammatic representation of the BONSAI ontology. The circles refer to the classes (and subclasses) defined in the BONSAI ontology. The text on the arrows defines the properties or predicates that connect the classes. All classes and predicates defined with the prefix "bont:" are terms specific to the BONSAI ontology. The classes defined in green circles are classes specific to the BONSAI ontology. Grey circles refer to the OM ontology used to define units and measures; classes in the blue circles refer to the OWL time ontology to define temporal extent of data; class in yellow refers to Schema ontology used to define geographic location of the datapoint. The rectangular class refers to a literal used to define any datatype such as strings or numbers or dates. The green rectangle is a datapoint instance from a dataset annotated based on the ontology. The rounded rectangles refer to the metadata related to the datapoint. The predicates (using prefix "prov:") used to define provenance from the metadata are provided by the Prov-O ontology
*Note*: https://www.ontology-of-units-of-measure.org/
https://www.w3.org/ns/prov#

In YSTAFDB, the flow of a "reference material" or primary element/substance is mapped across different processes as the primary element transforms into products (e.g., refined metal, alloys, or tailing) defined under "material name." Hence, to annotate the *FlowObjects*, we combined the "reference material" and the "material name" into a single name. Similarly, we annotated *ActivityTypes* by combining "subsystem name" and "process name." Besides *FlowObjects* and *ActivityTypes*, we represent also the unit and direction (input or output) of flows . YSTAFDB includes extensive information on the source of the data for each flow. This was annotated using the Prov-O extension of the BONSAI ontology.

While the ontology was sufficient to annotate the EXIOBASE datasets, it could not be used to annotate additional information on criticality, and recyclability of materials from YSTAFDB. Including this information is not essential to perform LCSA. However, this information could be incorporated by developing extensions of the core ontology. For example, criticality and recyclability could be additional properties of *Flows*.

## 2.3 | Interoperability

To publish interoperable content in the Semantic Web requires that information providers agree upon common frameworks and common controlled vocabularies or ontologies for annotation. To enhance the possibility of interoperability, we used constructs established by RDFS, the OWL and Simple Knowledge Organization System (SKOS),[4] in combination with the RDF data model to define constraints that data must meet (Holland & Culture, 2010; Nath et al., 2017, 2020). Constructs defined using these ontologies help to define relationships to link distinct knowledge organization systems (Holland & Culture, 2010). We introduced interoperability at both ontological level and individual data level (see Figure 2).

[4] https://www.w3.org/2004/02/skos/core#

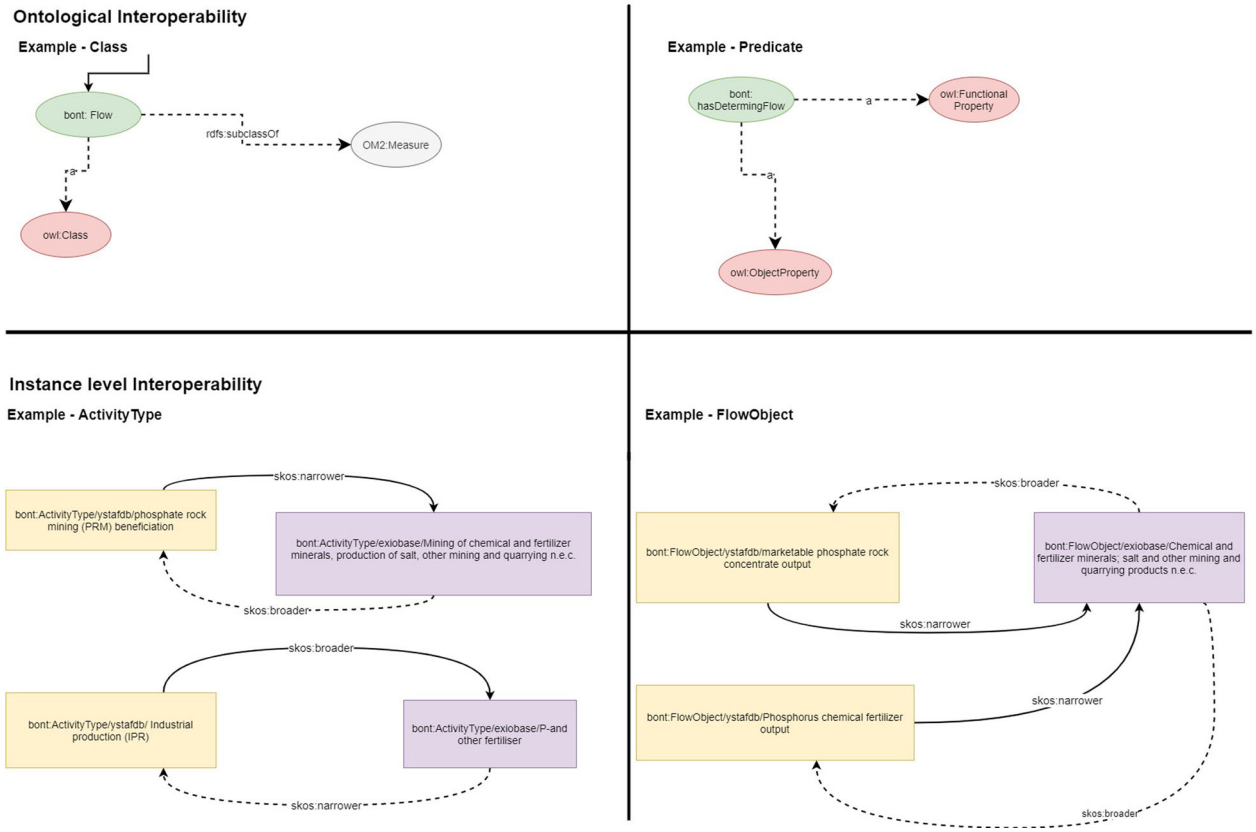


**FIGURE 2** Diagrammatic examples of using ontological and instance level interoperability. Each class in the BONSAI ontology is defined as a class by linking it to the class construct given in the OWL ontology; new predicates in BONSAI ontology are mainly defined as object properties. Specific properties such as bont:determiningFlow and bont:referenceFlowObject are further defined as functional properties, that is, connect to single unique instance IRI. In the example, for instance level interoperability, SKOS predicates skos:narrower and skos:broader were used to correspond the instances between EXIOBASE and YSTAFDB. The yellow rectangles refer to instances of ActivityTypes and FlowObjects from YSTAFDB. The violet rectangles refer to instances of ActivityTypes and FlowObjects from EXIOBASE

### 2.3.1 | Ontological interoperability

Ontological interoperability refers to linking of the core ontology to existing semantic ontologies. For example, each new "Class" in the BONSAI ontology refers to the construct "owl:Class" as defined by the OWL ontology. If a specific new class has the attributes of another class defined in an external ontology it can be linked using the predicate "rdfs:subClassOf" described by RDFS. As explained in an earlier example the class "Flow" defined in the BONSAI ontology is a measure of an entity, hence it is linked to the class "Measure" as defined by the OM2 ontology.

Besides classes, new predicates defined in the BONSAI ontology were also linked to the OWL ontology to better define their functions. For example, the predicate "bont: hasDeterminingFlow" was defined as an "owl: ObjectProperty" and a "owl:FunctionalProperty." Object property refers to predicates that link two different IRIs. A predicates that links an IRI to a literal (e.g., a string, number, date, time) is further defined to have a Datatype property. A functional property is a property that can determine that the predicate can link an IRI of an instance to only one other (unique) IRI of an instance.

### 2.3.2 | Instance level interoperability

With respect to interoperability of the datasets, we tested the similarities between EXIOBASE v3 and YSTAFDB. When describing the correspondence between the Activity types in EXIOBASE and YSTAFDB, we identified an instance of an activity type in YSTAFDB did not exactly correspond to an instance of an activity type in EXIOBASE. Instances of activity types and flow objects in YSTAFDB corresponded either broadly or narrowly to instances of an activity types and flow objects in EXIOBASE. Such relationships can be semantically described using mapping predicates such as *skos:broader*, *skos:narrower*, *skos:closeMatch*, and *skos:exactMatch* given by the SKOS ontology.

The data modeled in EXIOBASE is from 2011, while relevant data from YSTAFDB for this specific year was available only for flows and activities related to phosphorus. The instance level interoperability was initially done manually and further linked using the SKOS ontology. We used SKOS predicates *skos:broader* and *skos:narrower* to link the flows and activities related to phosphorus in YSTAFDB to flows and activities for fertilizer production and use in EXIOBASE (See Figure 2). These predicates have inverse property, that is,

<A> *skos:narrower* <B>.

Automatically entails

<B> *skos: broader* <A>.

## 3 | DATA STORAGE AND USE

As a result of the extraction process we generated a database containing over 230 million RDF triples from the EXIOBASE and the YSTAFDB database. All data extracted as triples was stored in OpenLink Virtuoso Triplestore. Data stored in triplestores is retrieved using SPARQL query language. We set-up the BONSAI SPARQL endpoint to query the datasets annotated using the BONSAI ontology. The endpoint was developed using yasgui query interface (TriplyDB, 2020).

The applicability of the ontology to the datasets is based on its ability to extract the required information. To do so, domain experts develop competency questions. Domain experts, in our case, were members of the BONSAI community who actively work with different tools and databases used in industrial ecology.[5] We asked the experts how they use the specific databases and the kind of data information they would like to extract from the databases. This information was used to develop the competency questions. These were further translated into SPARQL queries to extract the required information from the database (see Table 3 for examples). Thus, competency questions allow to verify the appropriateness of the model against the stored data. More example queries are available at the BONSAI SPARQL endpoint including the possibility of users writing their own queries.[6,7]

## 4 | DISCUSSION

In this section, we discuss the general validity of the proposed ontology, the opportunities and challenges of linking data to the Semantic Web, and suggestions for future work.

### 4.1 | General applicability of the proposed ontology

The ontology proposed here is intended for practical application in LCSA. Core datasets in EXIOBASE and YSTAFDB were annotated using the ontology and data from these sources can be independently extracted.[7,8] This has been studied by addressing the competency questions as discussed above. Table 3 gives examples of some competency questions and their corresponding queries. Additional examples can be found in www.odas.aau.dk

Table 4 provides a comparison of the BONSAI ontology with existing ontologies used to structure data for sustainability assessment. The BONSAI community resolved to build on prior ontologies used to define primary concepts in environmental sustainability assessment, particularly with respect to LCA. LCA is a widely used sustainability assessment tool and provides the conceptual framework for LCSA (Troullaki et al., 2021). Existing LCA ontologies define the key elements of life cycle inventory and LCA data. However, to be applicable to all kinds of LCSA datasets it was important to include broader definitions within the primary classes. As recommended by Pauliuk et al. (2016), the BONSAI ontology avoids rigid definitions of concepts such as classifying terms based on their extrinsic properties (e.g., classifying a flow as a resource, waste, or emission) or determining a system boundary between the technosphere and the natural system. For example, the class of ***FlowObjects*** were given a very broad definition that encompasses natural and manufactured physical assets, intellectual, human and social network assets, as well as financial assets. Similarly, the class of Activities are broadly defined, encompassing both natural and human-controlled activities, including production, consumption, and trade activities, as well as stock change activities (stock additions and stock removals). The ontology as proposed can thus be used to annotate data from socioeconomic datasets into the Semantic Web. For example, data from World bank on GDP per capita could also be annotated using BONSAI ontology. All economic activities can be defined as ***ActivityTypes***; labor costs, net taxes, net operating surplus (***FlowObjects***) can be defined as input or output ***Flows***; population (such as households, individuals, persons, legal persons) can be defined as ***Agents***.

[5] https://github.com/orgs/BONSAMURAIS/people
[6] https://odas.aau.dk
[7] https://github.com/BONSAMURAIS/yasgui-query-interfa-ce

**TABLE 3** Competency questions (and their corresponding SPARQL queries) used to evaluate the ontology with respect to the datasets converted used for data annotation to RDF

| Competency question | SPARQL query<br>Check BONSAI SPARQL endpoint (http://odas.aau.dk) for complete list of queries and live results | Result |
|---|---|---|
| What are the different *ActivityTypes* in a given dataset? | PREFIX bont: http://ontology.bonsai.uno/core#<br>PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#<br>PREFIX om2: http://www.ontology-of-units-of-measure.org/resource/om-2/<br>PREFIX btime: http://rdf.bonsai.uno/time#<br>PREFIX prov: http://www.w3.org/ns/prov#<br><br>SELECT *<br>FROM <http://rdf.bonsai.uno/data/exiobase3_3_17/hsup><br>FROM <http://rdf.bonsai.uno/data/exiobase3_3_17/huse><br>FROM <http://rdf.bonsai.uno/data/exiobase3_3_17/emission><br>FROM <http://rdf.bonsai.uno/location/exiobase3_3_17><br>FROM <http://rdf.bonsai.uno/flowobject/exiobase3_3_17><br>FROM <http://rdf.bonsai.uno/activitytype/exiobase3_3_17><br>FROM <http://rdf.bonsai.uno/location/ystafdb><br>FROM <http://rdf.bonsai.uno/flowobject/ystafdb><br>FROM <http://rdf.bonsai.uno/activitytype/ystafdb><br>FROM <http://rdf.bonsai.uno/unit><br>FROM <http://rdf.bonsai.uno/time><br><br>WHERE<br>{<br>?activity a bont:ActivityType .<br>?activity rdfs:label ?label<br>} | Provides the IRI and label for each activityType in given dataset<br>EXIOBASE : 164<br>YSTAFDB: 2190 |
| What are the different *FlowObjects* in a given dataset? | PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT *<br>FROM <Specify IRI of dataset, See query 1><br>WHERE<br>{<br>?FlowObject a bont:FlowObject .<br>?FlowObject rdfs:label ?label<br>} | Provides the IRI and label for each *FlowObject* in given dataset.<br>EXIOBASE: 272 (200 products and 72 emission types)<br>YSTAFDB: 670 |
| In which unit are the *FlowObjects* measured? | **Example: List the FlowObjects in the YSTAFDB dataset and the corresponding units.**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT DISTINCT ?FlowObject ?unitLabel<br>FROM <Specify IRI of dataset, See query 1><br>WHERE<br>{<br>?x a bont:FlowObject;<br>rdfs:label ?FlowObject .<br><br>?z a bont:Flow; bont:hasObjectType ?x;<br>om2:hasUnit ?unit .<br><br>?unit rdfs:label ?unitLabel<br>} | Provides the labels of 670 *FlowObjects* in YSTAFDB and the corresponding unit it is measured in |
| Activities where a specific *FlowObject* is an output | **Example: Which economic activities emit "Carbon dioxide, fossil" and how much ?**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT ?activityType ?location (xsd:string(?value) as ?value) ?unit<br>FROM <Specify IRI of dataset, See query 1><br><br>WHERE<br>{<br>?flow a bont:Flow .<br>?flow bont:isOutputOf ?act .<br>?act bont:hasLocation / rdfs:label ?location .<br>?act bont:hasActivityType / rdfs:label ?activityType .<br>?flow bont:hasObjectType / rdfs:label "Carbon dioxide, fossil" .<br>?flow om2:hasNumericalValue ?value .<br>?flow om2:hasUnit / rdfs:label ?unit .<br>} | Provides the label of the ActivityType, its location, value/amount of *FlowObject* (carbon dioxide, fossil) and unit |

(Continues)

**TABLE 3** (Continued)

| Competency question | SPARQL query<br>Check BONSAI SPARQL endpoint (http://odas.aau.dk) for complete list of queries and live results | Result |
|---|---|---|
| What are the output flows from an Activity in a given location? | **Example: Query on EXIOBASE Cultivation of wheat in Denmark**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT ?FlowObject (xsd:string(?value) as ?value) ?unit<br>FROM <Specify IRI of dataset, See query 1><br>WHERE<br>{<br>{<br>?flow a bont:Flow .<br>?flow bont:isOutputOf ?act .<br>?act bont:hasLocation / rdfs:label "DK".<br>?act bont:hasActivityType / rdfs:label "Cultivation of wheat".<br>?flow bont:hasObjectType / rdfs:label ?FlowObject .<br>?flow om2:hasNumericalValue ?value .<br>?flow om2:hasUnit / rdfs:label ?unit .<br>} | The query results in 33 flows.<br>Provides the total amount of wheat produced in Denmark, including environmental emissions linked directly to this activity |
| What are the input flows to an Activity in a given time and location? | **Example: Query on EXIOBASE for "Cattle farming" in Australia**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT ?FlowObject (xsd:string(?value) as ?value) ?unit<br>FROM <Specify IRI of dataset, See query 1><br><br>WHERE<br>{<br>{<br>?flow a bont:Flow .<br>?flow bont:isInputOf ?act .<br>?act bont:hasLocation / rdfs:label "AU".<br>?act bont:hasActivityType / rdfs:label "Cattle farming".<br>?act bont:hasTemporalExtent btime:2011.<br>?flow bont:hasObjectType / rdfs:label ?FlowObject .<br>?flow om2:hasNumericalValue ?value .<br>?flow om2:hasUnit / rdfs:label ?unit .<br>} | The query results in 1599 flows.<br>Provides the label and total amount of products (*FlowObject*) required to produce Cattle in Australia |
| Is a specific *FlowObject* an output from a specific Activity? | **Example: Query on YSTAFDB if "Phosphorus in Food waste" an output of "Food processing"**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT (xsd:string(?z) as ?isRequired)<br>FROM <Specify IRI of dataset, See query 1><br>WHERE<br>{<br>bind (exists {<br>?xFlow bont:isInputOf ?yActivity;<br>^bont:hasDeterminingFlow ?yActivity;<br>bont:hasObjectType ?xObject .<br><br>?xObject rdfs:label "food waste output;P" .<br>?yActivity bont:hasActivityType / rdfs:label "Agricultural production;food processing" .<br>} as ?z)<br>} | The query results in a binary response of True or False if the *FlowObject* is an input or not for a specific activity |
| What is the determining flow for specific activity? | **Example: Query on EXIOBASE to identify determining flow for "Copper production"**<br>PREFIX <Specify namespace of ontology used. See query 1><br><br>SELECT ?FlowObject ?activityType<br><br>FROM <Specify IRI of dataset, See query 1><br><br>WHERE<br>{<br>?f a bont:Flow;<br>bont:hasObjectType ?fobject .<br><br>?fobject rdfs:label ?FlowObject . | The query results in the label of the *FlowObject* "Copper products" which is the determining flow of the Activity "Copper production" |

(Continues)

**TABLE 3** (Continued)

| Competency question | SPARQL query<br>Check BONSAI SPARQL endpoint (http://odas.aau.dk) for complete list of queries and live results | Result |
|---|---|---|
| | ?a a bont:Activity;<br>bont:hasDeterminingFlow ?f;<br>bont:hasLocation ?l;<br>bont:hasTemporalExtent btime:2011;<br>bont:hasActivityType / rdfs:label "Copper products" .<br>} | |
| What is the provenance of a specific flow in a given database? | **Example: Query on YSTAFDB to identify the provenance for the flow for Copper production "production, Cu" in Australia**<br>PREFIX <Specify namespace of ontology used. See query 1><br>DESCRIBE ?dataset<br>FROM <Specify IRI of dataset, See query 1><br>WHERE<br>{<br>?dataset<br>prov:hadMember <http://rdf.bonsai.uno/data/ystafdb/huse#F_25184>.<br>filter (contains(xsd:string(?dataset), "dataset")) .<br>} | The query results in the label of the source title "Copper and zinc recycling in Australia - potential quantities and policy options" and its identifier doi "10.1016/j.jclepro.2006.06.023" |

The basic data structure proposed in the ontology with *FlowObject*, *Flow*, *ActivityType*, and *Activity* is also meant to be relevant for other tools central to the domain of Industrial Ecology, notably Material Flow Analysis (MFA) and Agent-Based Modeling (ABM). For the purposes of testing the BONSAI ontology, information that are not strictly relevant for LCSA, for example, information on criticality and recyclability of materials from YSTAFDB, was not currently extracted. Similarly, although the ontology has a class for *Agent*, datasets for ABM often have additional information on behavioral parameters. There may also be a need to specify more precisely such parameters that an Agent can have relative to an Activity. Such information could be extracted and stored as triples unrelated to the BONSAI ontology, if desired by developing expansions of the core ontology. By addressing such requirements, expanding of our proposed LCSA ontology with these MFA and ABM-relevant concepts would complete the ontology from an Industrial Ecology perspective. These concepts may be included in future expansions of the BONSAI ontology.

In general, there is a lack of a common platform to exchange data within industrial ecology (Pauliuk et al., 2016, 2019). Variations in data structure of common datasets used for sustainability adds to this challenge. To ensure transparency and interoperability of datasets, earlier studies suggested the possibility to link the ontologies to the Semantic Web (Janowicz et al., 2015; Pauliuk et al., 2019). Although relational databases are easy to set up and maintain now, in future there will be challenges when we have silos of data. For example, it is difficult to frequently update the information and manage large repositories in conditions when data need to be continuously collected (Sobhkhiz et al., 2021). Moreover, the possibility of interoperability and linking data across and beyond use in industrial ecology models can be achieved using the LOD platform. Use of semantics provide reasoning opportunities and interpretive support. This is especially useful when user requirements are qualitative (Sobhkhiz et al., 2021). In this study, we explicitly linked the proposed ontology with the Semantic Web, including linking it to other established ontologies. For example, mapping provenance in the ontology increases transparency. Using the Prov-O ontology was useful to annotate the elaborate metadata particularly provided in YSTAFDB that includes source for each datapoint. Similarly, using SKOS ontology was useful to annotate interoperability between the two datasets.

It is vital to note that one semantic ontology cannot capture all aspects of a domain. Ontology developers in the life sciences have adopted the application ontology approach (Kamdar et al., 2019; Yeumo et al., 2017). Using this approach, developers concentrate on mapping and reuse of existing ontologies further developing it to suit a given application.

## 4.2 | Challenges and opportunities

The Global LCA Data Access Network (GLAD) has recognized the need to enhance accessibility and interoperability of LCA datasets (GLAD, 2020). Currently, while LCSA- relevant data exist on the Web, such data is fragmented in isolated sources (e.g., in relational databases, or flat files such as spreadsheets) not always openly accessible. Earlier studies that have developed ontologies for LCSA mentioned the "potential" of Semantic Web to overcome data integration challenges (Hertwich et al., 2018; Kuczenski et al., 2016; Pauliuk et al., 2016). Semantic processing and data integration

**TABLE 4** Comparison of the proposed BONSAI ontology with existing ontologies to structure data for sustainability assessments

| BONSAI ontology classes | General data model (Pauliuk et al., 2019) | LCA ontology (Kuczenski et al., 2016) | LCA ontology (Janowicz et al., 2015) | Description |
|---|---|---|---|---|
| bont:Activity | Process; Stock | Activity | Activity | Transformation, distribution, storage processes |
| bont:Flow | Flow | Exchange | Flow | Input or output of an entity to or from an Activity |
| bont:FlowObject | Object | Flow* | Elementary Flow; Intermediate Flow; Product | Objects of interest (goods, substances, commodities, materials, products, waste, environmental flows) |
| Schema:Place | Location | Spatial Scope | Location (*GeoSPARQL*)** | Geographic location |
| Time: ProperInterval; Time : Instant; xsd:date | Time | Temporal Scope | Time (*OWL*)** | Location in time (historic time, future [model] time, time point, time interval) |
| Agent | — | — | Agent | Individual or collective entities such as organizations or groups (households, governments, capital ownership, etc.) |
| OM: Measure; bont:BalanceableProperty | Layer | FlowQuantity | Unit of measure (*QUDT*) ** | Unit of measurement (mass, volume, economic value) |
| bont:ReferenceUnit | General ratio (property of a flow) | — | — | Measure to which all flow measures are proportional to (per CO2 eq., per capita) |

*Must include flow compartment to define if the entity belongs to biosphere or technosphere.
**Suggested linking to other ontologies on the Semantic Web but not implemented.

is the methodology through which researchers can query, retrieve, integrate, and analyze data and knowledge from multiple sources on the Web without the requirement on the part of the researchers to download and manually integrate those sources (Kamdar et al., 2019). The possibility of linking data to the Semantic Web using a common ontology improves the findability, accessibility, interoperability, and reproducibility of data. However, the uptake of Semantic Web modeling, which could overcome these limitation, in the LCSA domain has been limited. In this section we discuss the challenges and potential solutions for semantic processing.

### 4.2.1 | Usability

Currently there is a steep learning curve to understand and use Linked Data and Semantic Web technologies for researchers not accustomed to these models. Researchers need extensive programming skills to convert, integrate, query, and explore the data and knowledge sources. For example, in this current study we use RDFLib and SPARQL queries to integrate and retrieve information, respectively, from the datasets converted to machine-readable format using BONSAI ontology. Currently, developing sophisticated SPARQL queries is a highly technical process and possesses as a high cognitive entry barrier (Kamdar et al., 2019).

The data conversion software[7,8] developed in this study provides an initial template for researchers interested in integrating data with the Semantic Web. Similarly, we have developed multiple example queries to help researchers interested in extracting data as well as provide a template to develop new queries. Several auxiliary files are produced during the mapping as datasets are converted from excel or .csv to RDF strings. This expands the need for disk space in order to include additional information. However, this is a minor problem as disk space is becoming cheaper. In future, there is a need to develop web applications and visualizations to automate the work process and make it easy for LCSA researchers to integrate, query, and explore data across the Semantic Web. It is important to note that, once data is stored and represented as LOD, a new set of advanced data discovery opportunities arise particularly with respect to data accessibility and interoperability. The extensive use of Semantic Web expands the possibility of data discoverability and accessibility beyond a single research domain.

### 4.2.2 | Interoperability

One of the challenges in adopting the ontology is the interpretation of the nomenclature from different data sources. Flows or activities are classified or grouped differently in different data sources. For instance, "Mining of chemical and fertilizer minerals" is an aggregated activity in EXIOBASE, while in YSTAFDB such activities are disaggregated according to the type of fertilizer produced (e.g., phosphate rock mining for phosphorus-based fertilizer). Using existing W3C ontologies such as SKOS allows datasets to be connected not in only one-to-one linear relationship but also by more complex relationships such as one-to-many or many-to-one (Morales & Orrell, 2017). An opportunity currently considered for future work is to link the ontology with datasets beyond the industrial ecology domain such as natural sciences or geospatial data as these could potentially add value to the current datasets (Maus et al., 2020). Furthermore, interoperability could be improved by linking international classification systems with the BONSAI ontology such as the Classification of Economic Activities in the European Community (NACE) and the Harmonized Commodity Description and Coding Systems (HS classification). This is because international datasets are often developed in accordance to these international classification systems. When such information is readily available, linking these classifications with the ontology can be beneficial to achieve instance level data instead of developing a new correspondence between datasets manually.

### 4.2.3 | Semantic heterogeneity

This refers to the lack of reuse of existing semantic ontologies. This results from independent creation and evolution of multiple autonomous ontologies that are tailored to the requirements of a specific domain and application (Kamdar et al., 2019). Linked data principles emphasize the correct reuse of existing vocabularies as well as linking to entities that already exist on the Semantic Web using their IRIs (Bizer et al., 2011). In the BONSAI ontology we have used broad terminologies which align with previous ontologies suggested in the LCSA domain. This provides the opportunity to expand the ontology based on the requirements of other industrial ecology applications ensuring a linked data cloud. The BONSAI ontology itself is linked to multiple established and commonly used ontologies in the Semantic Web such as Schema, OM2, and Prov-O.

Currently, the BONSAI ontology is used to convert open data from various sources and in different formats into a single platform using a common conceptual structure. Once the data is converted, relevant data can be extracted according to a user's need. This corresponds to a form of data warehousing where all data are transformed under a common schema (Kamdar et al., 2019). This overcomes the semantic heterogeneity problem (Williams et al., 2012). Data cleaning, preservation, and easier indexing and querying are other common advantages of data warehousing.

### 4.3 | Future work

On a broader perspective this work takes the initial step forward toward the larger vision of an industrial ecology based on open data. In particular, the development of a functional ontology plays a central role for the BONSAI initiative, which aims at developing an evolving open database for LCSA, where institutions and organizations can actively contribute with data, algorithms, and user interfaces. The overall architecture of such an open database includes four main elements.

The first element is the harvesting and parsing of data from different sources: existing databases for LCA and IO data, other large data sources, and user-provided data. The second element of the architecture is the core database, where the ontology developed here is a key element because it allows a meaningful linking and storing of the diverse harvested data under a common model. The data collected needs to be validated (e.g., by checking compliance to the schema and the ontology) and quality checked via specific review processes while taking into account the related uncertainties. This constitutes the third element of the architecture. The fourth element of the architecture concerns the processing and use of the data in context, that is, for the practical purpose of performing LCSAs. In particular, this includes procedures for accessing the data, combining activity datasets into system models, data visualization and communication, and procedures regulating how the community can take part in the management and further development of the database. This study covers the first two elements of the overall architecture while the other two are objectives to be developed in future work.

## 5 | CONCLUSION

Effective and transparent sustainability assessment requires access to data from a variety of heterogeneous sources across countries, scientific and economic sectors, and institutions. We have proposed an ontology capable of modeling processes describing product life cycles. Using this ontology, we were able to link datasets to the Semantic Web providing a suite of software components and queries able to support data integration and extraction. We believe that this effort reduces the low barrier for cross-dataset analysis.

In conclusion, we wish to stress the importance of the community effort needed to derive and maintain such an ontology. Previous attempts to develop ontologies in the domain of LCA have yielded conflicting results and stalled at a seminal stage due to either scientific or practical reasons such as lack of testing, implementation tools, funding, or community support. Therefore, we believe this kind of ontology development must be supported by a robust community, ideally experts from different scientific disciplines in the area of engineering and environment, as well as computer science, with a common interest in industrial ecology.

### APPENDIX

Instructions and code required for reproducing the work presented in this study can be accessed at https://github.com/BONSAMURAIS/Bonsai_ODA_Reproducibility

### ACKNOWLEDGMENTS

The authors are thankful to members of BONSAI—Chris Mutel, Tomás Navarrete Gutiérrez, Miguel F. Astudillo, Michele De Rosa, Stefano Merciai, Arthur Jakobs, Søren Løkke, Katja Hose, and Massimo Pizzol for constructive technical solutions and comments for this research.

### CONFLICT OF INTEREST

The authors declare no conflict of interest

### DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created in this study. However, the data used to assess the applicability of the ontology are available at http://odas.aau.dk/. To ensure reproducibility of the proposed work, we have uploaded the necessary code and instructions at https://github.com/BONSAMURAIS/Bonsai_ODA_Reproducibility.

### ORCID

*Agneta Ghose* https://orcid.org/0000-0003-1972-1433
*Matteo Lissandrini* https://orcid.org/0000-0001-7922-5998
*Emil Riis Hansen* https://orcid.org/0000-0003-4103-1244
*Bo Pedersen Weidema* https://orcid.org/0000-0003-1863-6528

## REFERENCES


Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S. & Zhao, J. (2013). PROV-O: The PROV ontology. https://www.w3.org/TR/prov-o/#:~:text=Introduction

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43.

Bertin, B., Scuturici, V.-M., Pinon, J.-M. & Risler, E. (2012). CarbonDB: A semantic life cycle inventory database. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2683–2685). https://doi.org/10.1145/2396761.2398725

Beylot, A., Secchi, M., Cerutti, A., Merciai, S., Schmidt, J., & Sala, S. (2019). Assessing the environmental impacts of EU consumption at macro-scale. *Journal of Cleaner Production*, *216*, 382–393. https://doi.org/10.1016/j.jclepro.2019.01.134

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In A. Sheth (Ed.), *Semantic services, interoperability and web applications: Emerging concepts* (pp. 205–227). IGI Global. https://doi.org/10.4018/978-1-60960-593-3.ch008

Brickley, D., & Guha, R. (2014). *RDF schema 1.1*. https://www.w3.org/TR/rdf-schema/

Cooper, J., Noon, M., Jones, C., Kahn, E., & Arbuckle, P. (2013). Big data in life cycle assessment. *Journal of Industrial Ecology*, *17*(6), 796–799. https://doi.org/10.1111/jiec.12069

Crawford, R. H., Bontinck, P.-A., Stephan, A., Wiedmann, T., & Yu, M. (2018). Hybrid life cycle inventory methods – A review. *Journal of Cleaner Production*, *172*, 1273–1288. https://doi.org/10.1016/j.jclepro.2017.10.176

Davis, C., Nikolic, I., & Dijkema, G. P. J. (2010). Industrial ecology 2.0. *Journal of Industrial Ecology*, *14*(5), 707–726. https://doi.org/10.1111/j.1530-9290.2010.00281.x

DBpedia. (2019). *DBpedia*. https://wiki.dbpedia.org/

De Rosa, M., & Weidema, B. P. (2019). *BONSAI- Big Open Network for Sustainability Assessment Information*. https://bonsai.uno/

Exiobase Consortium. (2014). *Exiobase (v.3.3.17 hybrid)*. https://www.exiobase.eu/index.php/data-download/exiobase3hyb

Fathalla, S., Auer, S., & Lange, C. (2020). Towards the semantic formalization of science. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (pp. 2057–2059). Association for Computing Machinery. https://doi.org/10.1145/3341105.3374132

Ghali, M. R., & Frayret, J. M. (2019). Social semantic web framework for industrial synergies initiation. *Journal of Industrial Ecology*, *23*(3), 726–738. https://doi.org/10.1111/jiec.12814

GLAD. (2020). *Global LCA Access Data Network (GLAD)*. https://www.globallcadataaccess.org/about

Guinée, J. (2016). Life cycle sustainability assessment: What is It and what are its challenges?. In R. Clift & A. Druckman (Eds.), *Taking stock of industrial ecology* (pp. 45–68). Springer International Publishing. https://doi.org/10.1007/978-3-319-20571-7_3

Hansen, E. R., Lissandrini, M., Ghose, A., Løkke, S., Thomsen, C., & Hose, K. (2020). Transparent sharing and integration of life cycle sustainability data with provenance. In *International Semantic Web Conference*. Athens, Greece. https://iswc2020.semanticweb.org/

Hertwich, E., Heeren, N., Kuczenski, B., Majeau-Bettez, G., Myers, R. J., Pauliuk, S., Stadler, K., & Lifset, R. (2018). Nullius in verba1: Advancing data transparency in industrial ecology. *Journal of Industrial Ecology*, *22*(1), 6–17. https://doi.org/10.1111/jiec.12738

Hobbes, J., Pan, F., Cox, S., & Little, C. (2020). *Time ontology in OWL*. https://www.w3.org/TR/owl-time/

Holland, J., & Culture, M. (2010). Guidelines for mapping into SKOS, dealing with translations. https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ATHENA/Deliverables/D4.2_Guidelines%20for%20mapping%20into%20SKOS.pdf

Ingwersen, W. W. (2015). Test of US federal life cycle inventory data interoperability. *Journal of Cleaner Production*, *101*, 118–121.

Ingwersen, W. W., Hawkins, T. R., Transue, T. R., Meyer, D. E., Moore, G., Kahn, E., Arbuckle, P., Paulsen, H., & Norris, G. A. (2015). A new data architecture for advancing life cycle assessment. *The International Journal of Life Cycle Assessment*, *20*(4), 520–526.

Janowicz, K., Krisnadhi, A. A., Hu, Y., Suh, S., Weidema, P., Rivela, B., Tivander, J., Meyer, D., Berg-Cross, G., Hitzler, P., Ingwersen, W., Kuczenski, B., Vardeman, C., Ju, Y., & Cheatham, M. (2015). A minimal ontology pattern for life cycle assessment data. *Proceedings of the Workshop on Ontology and Semantic Web Patterns* (6th ed.). Citeseer.

Kamdar, M. R., Fernández, J. D., Polleres, A., Tudorache, T., & Musen, M. A. (2019). Enabling web-scale data integration in biomedicine through linked open data. *NPJ Digital Medicine*, *2*(1), 1–14.

Kuczenski, B., Davis, C. B., Rivela, B., & Janowicz, K. (2016). Semantic catalogs for life cycle assessment data. *Journal of Cleaner Production*, *137*, 1109–1117. https://doi.org/10.1016/j.jclepro.2016.07.216

Lavers Westin, A., Kalmykova, P., Rosado, L., Oliveira, F., Laurenti, R., & Rydberg, T (2019). Combining material flow analysis with life cycle assessment to identify environmental hotspots of urban consumption. *Journal of Cleaner Production*, *226*, 526–539.

Lissandrini, M., Mottin, D., Palpanas, T., Papadimitriou, D., & Velegrakis, Y. (2015). Unleashing the power of information graphs. *ACM SIGMOD Record*, *43*(4), 21–26. https://doi.org/10.1145/2737817.2737822

Mailk, A., McBain, D., Wiedmann, T. O., Lenzen, M., & Murray, J. (2019). Advancements in input-output models and indicators for consumption-based accounting. *Journal of Industrial Ecology*, *23*(2), 300–312. https://doi.org/10.1111/jiec.12771

Matentzoglu, N., Bail, S., & Parsia, B. (2013). A corpus of OWL DL ontologies. *Description Logics*, *1014*, 829–841.

Maus, V., Giljum, S., Gutschlhofer, J., da Silva, D. M., Probst, M., Gass, S. L. B., Lukeneder, S., Lieber, M., & McCallum, I. (2020). A global-scale data set of mining areas. *Scientific Data*, *7*(1), 289. https://doi.org/10.1038/s41597-020-00624-w

Merciai, S., & Schmidt, J. (2018). Methodology for the construction of global multi-regional hybrid supply and use tables for the EXIOBASE v3 database. *Journal of Industrial Ecology*, *22*(3), 516–531. https://doi.org/10.1111/jiec.12713

Mittal, V. K., Bailin, S. C., Gonzalez, M. A., Meyer, D. E., Barrett, W. M., & Smith, R. L. (2018). Toward automated inventory modeling in life cycle assessment: The utility of semantic data modeling to predict real-world chemical production. *ACS Sustainable Chemistry and Engineering*, *6*(2), 1961–1976. https://doi.org/10.1021/acssuschemeng.7b03379

Morales, L. G., & Orrell, T. (2017). Data interoperability: A practitioner's guide to joining up data in the development sector. https://www.data4sdgs.org/sites/default/files/services_files/Interoperability - A practitioner's guide to joining-up data in the development sector.pdf

Myers, R. J., Reck, B. K., & Graedel, T. E. (2019a). Yale stocks and flows database (YSTAFDB). https://doi.org/10.5281/ZENODO.2561882

Myers, R. J., Reck, B. K., & Graedel, T. E. (2019b). YSTAFDB, a unified database of material stocks and flows for sustainability science. *Scientific Data*, *6*(1), 84. https://doi.org/10.1038/s41597-019-0085-7

Nath, R. P., Hose, K., Pedersen, T. B., & Romero, O. (2017). SETL: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, *68*, 17–43. https://doi.org/10.1016/j.is.2017.01.005

Nath, R. P. D., Hose, K., Pedersen, T. B., Romero, O., & Bhattacharjee, A. (2020). SETLBI: An Integrated Platform for Semantic Business Intelligence. In *Companion Proceedings of the Web Conference 2020* (pp. 167–171). Association for Computing Machinery. https://doi.org/10.1145/3366424.3383533

Nolin, M.-A., Ansell, P., Belleau, F., Idehen, K., Rigault, P., Tourigny, N., Roe, P., Hogan, J. M., & Dumontier, M. (2008). Bio2RDF network of linked data. In Semantic Web Challenge; International Semantic Web Conference (ISWC 2008). Karlsruhe. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.210.3235&rep=rep1&type=pdf

Pauliuk, S., Heeren, N., Hasan, M. M., & Müller, D. B. (2019). A general data model for socioeconomic metabolism and its implementation in an industrial ecology data commons prototype. *Journal of Industrial Ecology*, *23*(5), 1016–1027. https://doi.org/10.1111/jiec.12890

Pauliuk, S., Majeau-Bettez, G., Müller, D. B., & Hertwich, E. G. (2016). Toward a practical ontology for socioeconomic metabolism. *Journal of Industrial Ecology*, *20*(6), 1260–1272. https://doi.org/10.1111/jiec.12386

Prud'hommeaux, E., & Seaborne, A. (2008). SPARQL query language for RDF. https://www.w3.org/TR/rdf-sparql-query/

Ragget, D. (2009). Introduction to linked data and Semantic Web technology. https://www.w3.org/2009/03/xbrl/talks/intro2semweb-dsr.pdf

Raimond, Y., Ramsden, D., Bartlett, O., & Angeletou, S. (2017). Linked data and the semantic web. https://www.bbc.co.uk/academy/en/articles/art20130724121658626

Rijgersberg, H., van Assem, M., & Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, *4*(1), 3–13.

Sala, S. (2020). Chapter 3 - Triple bottom line, sustainability and sustainability assessment, an overview. In J. Ren, A. Scipioni, A. Manzardo & H. Liang (Eds.), *Biofuels for a more sustainable future* (pp. 47–72). Elsevier. https://doi.org/10.1016/B978-0-12-815581-3.00003-8

Schema Community Group. (2011). Schema.org. https://schema.org/Place

Schmidt, J., & De Rosa, M. (2020). Certified palm oil reduces greenhouse gas emissions compared to non-certified. *Journal of Cleaner Production*, *277*, 124045. https://doi.org/10.1016/j.jclepro.2020.124045

Schmidt, J., Merciai, S., Munoz, I., De Rosa, M., & Astudillo, M. F. (2021). The Big Climate Database Version 1 - Methodology report (February), Version 1.0. http://denstoreklimadatabase.dk/

Sobhkhiz, S., Taghaddos, H., Rezvani, M., & Ramezanianpour, A. M. (2021). Utilization of semantic web technologies to improve BIM-LCA applications. *Automation in Construction*, *130*, 103842. https://doi.org/10.1016/j.autcon.2021.103842

Stevens, R., & Lord, P. (2009). Ontologies and life science data management. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 1960–1963). Springer. https://doi.org/10.1007/978-0-387-39940-9_631

Takhom, A., Suntisrivaraporn, B., & Supnithi, T. (2013). Ontology-enhanced life cycle assessment: A case study of application in oil refinery. In *The Second Asian Conference on Information Systems (ACIS), Phuket, Thailand*.

TriplyDB. (2020). Yasgui query interface. https://triply.cc/docs/yasgui

Troullaki, K., Rozakis, S., & Kostakis, V. (2021). Bridging barriers in sustainability research: A review from sustainability science to life cycle sustainability assessment. *Ecological Economics*, *184*, 107007. https://doi.org/10.1016/j.ecolecon.2021.107007

Weidema, B. P., Cappellaro, F., Carlson, R., Notten, P., Pålsson, A.-C., Patyk, A., Regalini, E., Sacchetto, F., & Scalbi, S. (2003). *Procedural guideline for collection, treatment, and quality documentation of LCA data*. Italian National Agency for New Technologies, Energy and the Environment. https://lca-net.com/files/V2004_ProceduralLCA.pdf

Weidema, B. P., Schmidt, J., Fantke, P., & Pauliuk, S. (2018). On the boundary between economy and environment in life cycle assessment. *The International Journal of Life Cycle Assessment*, *23*(9), 1839–1846. https://doi.org/10.1007/s11367-017-1398-4

Wick, M. (2012). About Geonames. http://www.geonames.org/about.html

Wick, M., & Vatant, B. (2012). The Geonames geographical database - GeoNames Ontology. https://www.geonames.org/ontology/documentation.html

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., …, & Mons, B. (2012). Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, *17*(21), 1188–1198. https://doi.org/10.1016/j.drudis.2012.05.016

Yeumo, E., Alaux, M., Arnaud, E., Aubin, S., Baumann, U., Buche, P., Cooper, L., Ćwiek-Kupczyńska, H., Davey, R. P., Fulss, R. A., Jonquet, C., Laporte, M.-A., Larmande, P., Pommier, C., Protonotarios, V., Reverte, C., Shrestha, R., Subirats, I., Venkatesan, A., Whan, A., & Quesneville, H. (2017). Developing data interoperability using standards: A wheat community use case. *F1000Research*, *6*, 1843. https://doi.org/10.12688/f1000research.12234.1

Zhang, Y., Luo, X., Buis, J. J., & Sutherland, J. W. (2015). LCA-oriented semantic representation for the product life cycle. *Journal of Cleaner Production*, *86*, 146–162.

**How to cite this article:** Ghose A, Lissandrini M, Hansen ER, Weidema BP. A core ontology for modeling life cycle sustainability assessment on the Semantic Web. *J Ind Ecol*. 2021;1–17. https://doi.org/10.1111/jiec.13220