Transparent Integration and Sharing of Life Cycle Sustainability Data with Provenance

Hansen, Emil Riis; Lissandrini, Matteo; Ghose, Agneta; Løkke, Søren; Thomsen, Christian; Hose, Katja

Published in:
The Semantic Web – ISWC 2020 - 19th International Semantic Web Conference, 2020, Proceedings

DOI (link to publication from Publisher):
10.1007/978-3-030-62466-8_24

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Transparent Integration and Sharing of Life Cycle Sustainability Data with Provenance

Emil Riis Hansen[1(✉)], Matteo Lissandrini[1], Agneta Ghose[2], Søren Løkke[2], Christian Thomsen[1], and Katja Hose[1]

[1] Department of Computer Science, Aalborg University, Aalborg, Denmark
{emilrh,matteo,chr,khose}@cs.aau.dk
[2] Department of Planning, Aalborg University, Aalborg, Denmark
{agneta,loekke}@plan.aau.dk

Abstract. Life Cycle Sustainability Analysis (LCSA) studies the complex processes describing product life cycles and their impact on the environment, economy, and society. Effective and transparent sustainability assessment requires access to data from a variety of heterogeneous sources across countries, scientific and ecsonomic sectors, and institutions. Moreover, given their important role for governments and policymakers, the results of many different steps of this analysis should be made freely available, alongside the information about how they have been computed in order to ensure accountability. In this paper, we describe how Semantic Web technologies in general and PROV-O in particular, are used to enable transparent sharing and integration of datasets for LCSA. We describe the challenges we encountered in helping a community of domain experts with no prior expertise in Semantic Web technologies to fully overcome the limitations of their current practice in integrating and sharing open data. This resulted in the first nucleus of an open data repository of information about global production. Furthermore, we describe how we enable domain experts to track the provenance of particular pieces of information that are crucial in higher-level analysis.

Keywords: Open data · Provenance · Sustainability analysis

1 Introduction

Sustainability is increasingly becoming a key aspect both for policy making and commercial positioning. Its importance is expected to increase with the global socioeconomic impacts of climate change [11,15]. Life Cycle Sustainability Analysis (LCSA) studies the impacts of products along their life cycle, from the extraction of raw materials to their production, and till their disposal [3]. This enables enterprises and organizations to assess the impact of their current production chain and to find more sustainable means of production, also in line

J. Z. Pan et al. (Eds.): ISWC 2020, LNCS 12507, pp. 378–394, 2020.
https://doi.org/10.1007/978-3-030-62466-8_24

with the goal of sustainable development [15]. Despite this crucial role, large variations in assumptions and origins of data embedded in the assessments hinder the reliability of the outcome of such analyses. Given the complex nature of the production chain of any product, to perform reliable LCSA, analysts need access to data from a variety of heterogeneous sources across countries, scientific and economic sectors, and institutions. To enable the integration of diverse data sources, previous efforts [6] designed an ontology and corresponding open database to allow multiple organizations and researchers to share LCSA data and to make use of such data to produce analysis and models. These efforts lay the foundations of a platform where domain experts can both freely access data to compute and produce new models, but also re-share their results within the same framework. LCSA involves heterogeneous data-sources and actors, hence, it is important to assure transparency, verifiability, and reproducibility of the contents of any data involved in the process. This is achieved by tracking the provenance of the information employed. Information about provenance (also called lineage [16]) allows scientists to track their data through all transformations, analyses, and interpretations [1]. *In this work, we share our experience of opening up datasets from non-open formats.* This will both help any party interested in accessing and sharing LCSA data, as well as provide useful insights to any organization willing to publish their own data to foster open science.

Contributions: *This work presents an account of how Semantic Web (SW) technologies are "in use" in an Open Source Database for Product Life Cycle Sustainability Analysis*, in direct collaboration with domain experts and associations involved in open sustainability assessment (http://bonsai.uno). We first provide an introduction to the domain of Life Cycle Sustainability Analysis and its links with Semantic Web technologies (Sect. 2). We then describe how the BONSAI Open Database for Product Footprinting is tackling the problem of integrating heterogenous LCSA data within a single open knowledge base (Sect. 3). Further, we detail how we represent, keep track of, and allow querying for the provenance of each piece of information in our open data repository. In particular, we describe the data integration workflow and how this is supported by the current open LCSA ontology (Sect. 4). We then detail how the BONSAI Open Database allows modeling all the core data required to develop economic input-output models used in LCSA (Sect. 5). The workflow we implemented allows for integrating datasets from different sources and to republish them as Linked Open Data. Further, we explain how these datasets, once converted, are annotated with provenance information adopting the PROV-O [22] vocabulary, allowing to verify the lineage of the source data (Sect. 6). Finally, we present some important lessons learned while overcoming the challenges of employing SW technologies in this domain (Sect. 7).

2 Background and Domain

Data used to perform LCSA originates from multiple sources such as national statistics, environmental reports, and supply chain reports [20]. In addition, to

diversity in data sources, the data models also differ. For example, data on the production of goods or services can be reported in mass or monetary units. Therefore, to perform LCSA, domain experts need to integrate data sets from heterogeneous sources. Usually, LCSA relies on large databases (e.g., Ecoinvent[1]) that contain data at different levels of granularity about many human activities. Practitioners use these databases to compute specific models of the systems and processes they study. *Yet, many of these databases provide little to no access to the techniques used to collect and integrate the data. Moreover, in many cases, these databases are proprietary, expensive to access, and lack inter-operability.* Therefore, given the crucial role of LCSA, the BONSAI organization set out to overcome the current lack of accessibility and transparency with an Open Database for LCSA and developed, as a first step, an appropriate ontology [6]. *Here, we present how SW technologies have been adopted for the first time in LCSA to implement and materialize this open database.*

Availability and accessibility of up-to-date data, as well as legal and technical openness, are important elements that make Open Data the de-facto solution both in open science and in a more transparent government. In this spirit, other efforts have been taken in the direction of creating a database for LCSA analysis. The most notable are Exiobase, YSTAFDB, and Trase Earth. Exiobase [10] is a multi-regional Input-Output database that contains data on 200 product types that are transacted between 164 industries. Moreover, it contains records for 39 resource types, 5 land types, and 66 emission types related to the production and consumption of goods and services in the entire global economy. The Yale Stocks and Flows Database (YSTAFDB) combines material stocks and flows (STAF) data generated since early 2000 and collected by researchers at Yale [12]. Trase Earth[2] is another LCSA initiative that maps supply chain information system for land and forest use in Latin America.

Yet, while legal openness can be provided by applying an appropriate open license, technical openness requires us to ensure that there are no technical barriers to using the data. In particular, the aforementioned databases do not make use of Semantic Web technologies, which limits their ability to seamlessly integrate with other new datasets accessible on the Web. On the other hand, the success of many open-data resources in other domains such as GeoNames and Bio2RDF [13], motivates the decision to adopt Semantic Web technologies and the Linked Open Data format as a more appropriate solution. *Thus, while other efforts provided (legally) open datasets for product footprinting, this work is the first open database for LCSA on the Semantic Web.*

Nonetheless, while a common ontology and data format is the first step towards integrating and publishing free and open data for LCSA [6], *in this paper we focus on the next crucial step: integrating and sharing different Life Cycle Sustainability datasets.* Among others, we describe how we need not only to achieve full interoperability between different datasets, but also how we record and track the *data lifecycle* through provenance to ensure transparency,

[1] https://www.ecoinvent.org/.
[2] https://trase.earth/.

verifiability, and traceability of the original datasets and the computed results. To this end, we make use of the W3C PROV-O standard for modeling of provenance [22]. The standard has been widely used in different systems and contexts in the last couple of years. In general, the PROV-O vocabulary is highly flexible and enables the recording of lineage for any collection of data, recording activities (e.g., who gathered what, where, and when), which can be used to evaluate, among others, the reliability of the data. For instance, the W3C PROV-O standard has been used to expose provenance information regarding version control systems (VCS) [4], to enable the publication of VCS provenance on the Web and subsequent integration with other systems that make use of PROV-O. Moreover, it has been used for a Semantic Web-based representation of provenance concerning volunteered geographic information (VGI) [2] and to enhance the quality of an RDF Cube regarding European air quality [5].

3 Life Cycle Sustainability Data

Supply and Use Tables (SUTs) are one of the primary data sources for LCSA. They are comprehensive, non-proprietary data sources, covering the environmental, social, and economic spheres. In practice, the SUT records show what was the *total production* from a specific industrial sector and which other industrial sectors or markets consumed this product in which proportion. For instance, SUT records show that in 2011, ∼1237 megatonnes of *steel* were produced in *China* [10]. Furthermore, ∼92.7% of the domestic steel production in China was also used in China. Hence, a national SUT database encapsulates production and consumption of products and services for the entire national economy. Global Multi Regional (MR) SUTs are a combination of national SUTs and provide data on the global economy, which includes the transaction of goods and services between countries [10]. Among others, to measure the economic impacts of a change in demand of a specific product or service, Input-Output (IO) models are constructed from SUTs by applying one of the multiple algorithms existing in the literature [7]. IO models represent inter-industry relationships within an economy, showing how output from one industrial sector becomes an input to another industrial sector. As a result, an IO model, obtained from a set of MR SUTs, links flows of productions within and among national markets. If the SUTs include additional data on environmental emissions or social performance (e.g., employment levels), the IO models can be further used to perform environmental or social footprinting (e.g., the impact on carbon emissions by an increase in demand of a product).

A Model for Interoperable LCSA Data. Data available in multi-regional environmentally-extended IO models (EEIO) is aggregated for each industrial sector. Granularity in the analysis can be increased by combining the EEIO with detailed data of the product or service to be analyzed expanding it with different sources [17]. *However, what hinders this process is the lack of access to the relevant datasets and their limited interoperability.* To address this problem, we developed an Ontology for Product Footprinting to ease and promote the

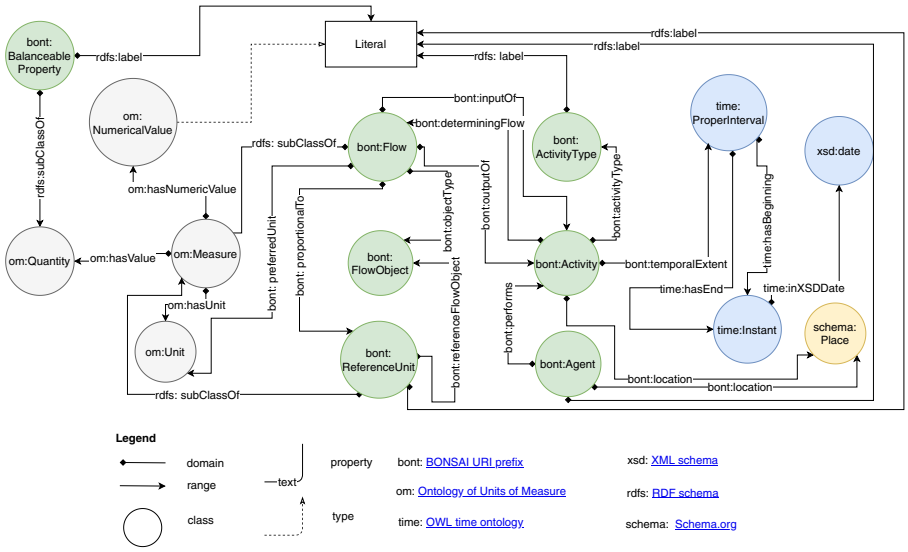

Fig. 1. The BONSAI ontology for LCSA [6]

exchange and integration of diverse LCSA data sources [6]. The proposed ontology (Fig. 1) follows a well-established model around three main concepts: *Activity* (any production activity, e.g., *steel production*), *Flow* (a quantity of product that is either produced or consumed by an activity, e.g., *tonnes of steel produced in China*), and *Flow Object* (the kind of product that is produced or consumed, e.g., *Stainless steel*) [21]. This ontology has been designed to model both economic production and environmental emissions. While the ontology presents a crucial first step for different stakeholders to agree on a common vocabulary and data model, additional tasks are required for the realization of a common open database. In the following, we describe such tasks. In particular, we have established a data integration workflow where multiple data sources are integrated to expand the granularity of the information and allow the construction of more detailed EEIO models. In the following, we adopt the Exiobase dataset and the YSTAFDB as prototypical example resources to demonstrate how we achieve the desired interoperability. *In particular, we describe how we enable integration and sharing of multiple SUTs within the common BONSAI Open Database.*

4 Data Integration Workflow

The integration workflow we established starts when a new dataset is identified for inclusion and terminates with the output of RDF named graphs representing the (annotated) information that was extracted from the identified dataset (see Fig. 2). The graphs are then published as Linked Open Data resources. We note that we tackle explicitly the task of integrating datasets with different formats

within a unique repository with a common data model. That is, we enforce (both manually and automatically) syntactic data quality, but we do not tackle the issue of fixing data quality issues in the content of the data we integrate. *This is on purpose since our goal is to collect and store multiple datasets as they are.* Inspecting and solving data quality issues is an orthogonal task that domain experts can carry out only when they have open access to different datasets to compare and cross-reference. This means that, without our open database, ensuring the quality of the data used in LCSA would not be easy (or not feasible at all). In Sect. 7, we provide an example of such a case.

Integration of Multiple Classifications. Different datasets might have distinct classifications for the same concept. To align those datasets, *correspondence tables* systematically encode the semantic correspondence between those concepts within the BONSAI classification. Correspondence tables, hence constitute a reference taxonomy being developed by BONSAI to keep track of conceptual linkage between various datasets. For example, the Exiobase dataset introduces 163 different instances of *Activity Types*, 200 *Flow Objects*, and 43 *Locations*. One of the instances is the *Activity Type* of *cultivation of paddy rice*. In this case, the new concept is added in the BONSAI classification (Fig. 2, top dashed arrow) recording that *cultivation of paddy rice* is an *Activity Type* in the BONSAI classification extracted from the Exiobase dataset.

Moreover, in Exiobase there is a special *Flow Object* labeled *"Other emissions"*. Within the BONSAI classification, this concept is also linked to a set of more specific emissions listed by the United States Environmental Protection Agency (US EPA). This correspondence is hence recorded via the `partOf` relation to make data within the two classifications interoperable. Establishing semantic equivalence requires some domain knowledge, hence correspondence tables are manually created. *Create Correspondence Table* is the first process in the data workflow (Fig. 2). Then we perform the process of *Correspondence Mapping*, which produces the new enhanced dataset containing the updated correspondence information (in Fig. 2 labeled *Correspondence Mapped Dataset*).

Intermediate Data Transformation. In the process of integrating new LCSA datasets, we faced the technical issue of many LCSA datasets being shared in various non-normative formats. As an example, the Exiobase dataset is shared as a set of spreadsheets, without an associated ontology. Similarly, YSTAFDB datasets are provided as plain CSV files. The data structure, even within the same file format (e.g., CSV files), might however also differ from dataset to dataset, due to lack of standardization between LCSA datasets [8]. To allow automatic transformation and integration of new datasets by a common set of data converters, we defined a common intermediate CSV format. The *Formalization Transformation* activity represents the conversion of the specific data-formats to the common one (in Fig. 2 with output *Formalized Dataset*). The formalized datasets will contain a separate list of *Flows*, *Flow Objects*, *Activity Types*, and *Locations*. Finally, this formalization task could also be carried out by any data provider who wants to include their dataset in the BONSAI database.

RDF Data Extraction. The final step in the integration of a new dataset is the actual conversion of the formalized data into an RDF graph coherent with the BONSAI Ontology. Custom scripts are used in this process (called *Data Extraction*) to create named graphs from the formalized data. The result is one or more named graphs with instances of *Flow Objects*, *Activity Types*, *Locations*, and *Flows* (*Named Graphs* in Fig. 2). Our convention is to create a named graph for each class of instances. Thus, if a new dataset presents *Locations*, *Flow Objects*, *Activities*, and *Activity Types* we create four new named graphs, one for each of the four classes. Furthermore, this convention tries to avoid duplicating concepts by storing them only once in their dedicated named graph. Since the same information usually appears in several datasets, the other datasets, when integrated, will just reference the information in the predefined named graph avoiding redundancies. Finally, the newly generated graphs can be published via a SPARQL endpoint. Moreover, while the BONSAI classification is expanded since new named graphs are produced and integrated in the database, the intermediate resources (in the dashed ovals) can be discarded. Finally, since the conversion script is automatic (due to the formalization step), we can ensure its conformity to the proposed ontology and also identify missing information. In our future work, we aim to also adopt shape expressions [14] for syntactic validation of extracted information.

Integration of new Models. After a new dataset is integrated and published, the database is used as a source of information to compute new or updated IO models. Development of IO models from MR SUTs varies depending on the algorithm used for IO Modeling [7,18]. Nonetheless, users of the BONSAI database can apply their own or predefined IO Modeling Algorithms to some or part of the data published in the database by querying only the required data. For instance, given that both Exiobase and YSTAFDB comply with the flow-activity model encoded in the ontology [6], data from both can be processed altogether or a user can select a portion of them for IO Modeling in a specific sector. This step is illustrated in Fig. 2 as the process *IO Modeling* using the named graphs in the database along with an IO Modeling Algorithm. The result of this process is a new named graph representing the *Flows* and the corresponding information in the IO model. This means that the database allows also the insertion of the IO models into the dataset (illustrated with a dashed line between the *IO Models* and the *Named Graphs*).

Metadata Annotation. For all systems that incorporate data from multiple diverse sources, keeping provenance information about individual pieces of data is crucial. For new datasets this corresponds to the information of their origin, especially the organization and the time at which they have been produced. For IO models this also includes the portion of the dataset used to compute them and the metadata about how they have been computed. Therefore, during the integration processes described above, the output datasets are also annotated with provenance information, as described in Sect. 5.

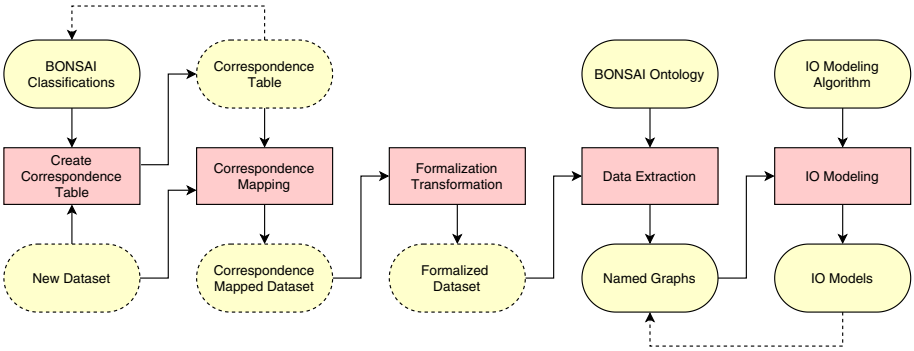

Fig. 2. Integration workflow of a new LCSA dataset. Squares represent processes, ovals represent data, arrows indicate the flow of data, dashed ovals represent data which is not saved after having been used in all their respective processes.

Handling Updates. The pipeline is rerun whenever a new dataset is integrated, or when a new version of an already integrated dataset is available. All steps of the pipeline must be rerun for the integration of new datasets, but changes to existing datasets often do not require the initial manual step of *Create Correspondence Table*, since the schema between versions of a dataset is rarely changed.

5 Support for Provenance

Provenance information is used to determine how an artifact was produced, and from where it origins. This allows, among others, to verify whether correct

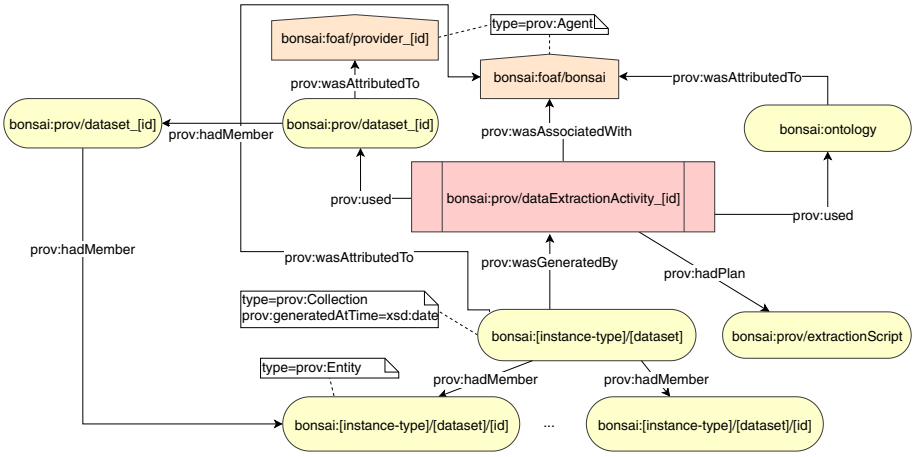

Fig. 3. Implementation of Provenance in the LCSA integration workflow. Pentagons, ellipses, and squares describe PROV-O *Agents*, *Entities*, and *Activities* respectively. Arrows represent PROV-O provenance relationships between model constituents.

```
bonsai:flowobject/exiobase_3_3_17 a prov:Collection ;
    prov:generatedAtTime "2019-11-28"^^xsd:date ;
    prov:wasGeneratedBy bonsai:prov/dataExtractionActivity_0 ;
    prov:hadMember bonsai:flowobject/C_ADDC,
        bonsai:flowobject/C_STEL,
        bonsai:flowobject/C_ALUM,
        ...
```

Fig. 4. Fragment of provenance record for the named graph for Exiobase v3.3.17 *Flow Objects*. Prefixes: `bonsai:` for BONSAI common resources (https://rdf.bonsai.uno/), and `prov:` for PROV-O (http://www.w3.org/ns/prov#)

methods have been utilized to obtain a result and hence whether artifacts can be trusted [22]. Thus, data in the BONSAI database is enhanced with provenance information annotations for all resources integrated through time.

In practice, the BONSAI provenance model is implemented by referencing and subclassing concepts from the W3C PROV-O vocabulary [9]. PROV-O uses the concepts of *Agents*, *Entities*, and *Activities*, to describe objects and their life cycle. In PROV-O *Entities* can be physical, digital, or conceptual objects of which we want to keep track. *Activities* are records of how entities come into existence and how existing *Entities* are changed to become new *Entities*. *Agents* can be a person, a piece of software, an organization, or other entities that may be ascribed responsibility [22]. Hence, PROV-O defines concepts to relate *Agents*, *Entities*, and *Activities* used in the production, delivery, or in other ways influencing an object [22]. Provenance information is automatically produced during the *Data Extraction* process in the data integration workflow described in Sect. 4, following the specific implementation as illustrated in Fig. 3. In the following, we explain how this is materialized using the integration of the Exiobase dataset as an example. At the time of writing, we have also integrated the YSTAFDB [12].

The integration of a new dataset results in the creation of one or more named graphs defining instances of *Flow Objects*, *Activity Types*, *Locations*, and *Flows*. Each named graph is assigned a unique URI (e.g., `bonsai:flowobject/exiobase_3_3_17`, for the named graph defining the *Flow Objects* extracted from Exiobase v3.3.17) and it is defined both as a distinct *Entity* and as a *Collection*. Also, a distinct URI is assigned to every instance (e.g., each instance of *Flow Object*, *Flow*, or *Activity Type*) in each *dataset* (e.g., *Exiobase*). For instance, in Exiobase (v3.3.17) the *Flow Object* describing *Basic iron and steel of ferro-alloys and first products thereof* (code `C_STEL`) has URI `bonsai:flowobject/exiobase_3_3_17/C_STEL`. Finally, the *Data Extraction* process encodes provenance information for the named graphs, linking each graph to both its data source and the version of the script used for data extraction. Moreover, it lists all the instances in the graph as members of the corresponding collection. That is, we record membership to a specific collection for each resource within each graph (lowest level of Fig. 3). In practice, the

PROV-O relation `prov:hadMember` is used to relate instances of data to the same `prov:collection`. This explicit link is materialized to improve accessibility to users and automatic analysis tools. The result of this model is an annotated resource associated with provenance information about the creation time of the named graph, which activity was used in its generation, and the list of its members. A fragment of a concrete example of such a record is shown in Fig. 4 (non-provenance metadata has been omitted from the figure for clarity).

As explained above, named graphs are created during the process *Data Extraction*. Since data extraction is a crucial activity for the creation of the named graphs, we encode information about this step using a PROV-O *Activity*. The *Activity* encodes information about what entities were used in the creation of the named graphs, which was associated with the activity, and references the actual implementation (e.g., the script used for data extraction). Each activity is assigned a unique URI (e.g., `bonsai:prov/dataExtractionActivity_[id]`). Hence, this record links the usage of a set of resources, along with a plan of execution, to a specific data extraction activity. A concrete example of such a record is illustrated in Fig. 5. The record shows how the BONSAI ontology and a dataset were used in the activity referred to as `dataExtractionActivity`, along with the plan `extractionScript`, linking to the version of the script used in the PROV-O *Activity*. As illustrated in Fig. 4, the PROV-O relation `prov:wasGeneratedBy` is used to relate the content of a named graph, to an extraction activity. Hence, we maintain a consistent link between individual instances of extracted data and their respective origin datasets.

```
bonsai:prov/dataExtractionActivity_0 a prov:Activity ;
    prov:hadPlan bonsai:prov/extractionScript ;
    prov:wasAssociatedWith bonsai:foaf/bonsai ;
    prov:used <http://ontology.bonsai.uno/core>,
        bonsai:prov/dataset_0 .
```

Fig. 5. Provenance record of a data extraction activity referring to the BONSAI ontology (`ontology.bonsai.uno/core`), Exiobase v3.3.17 (`bonsai:prov/dataset_0`), and the extraction script identified by `bonsai:prov/extractionScript`.

Finally, we record the specific data extraction activity (e.g., `bonsai:prov/dataExtractionActivity_0`) that extracted data from the specific dataset (e.g., `bonsai:prov/dataset_0`). Hence, each dataset integrated into the BONSAI database is given a unique URI (e.g., `bonsai:prov/dataset_0`, for the Exiobase dataset v3.3.17). Furthermore, the dataset provider (e.g., an organization, a government, or an individual) is also given a unique URI (e.g., `bonsai:foaf/provider_0`, for the Exiobase Consortium maintaining Exiobase). For instance, the provenance record for Exiobase dataset v3.3.17 is illustrated using the turtle format in Fig. 6. The record contains metadata about the dataset (e.g., the version 3.3.17), a link to the organization responsible for it

(e.g., Exiobase Consortium), and a date for the latest dataset update before integration into the BONSAI database (e.g., 2019-03-12).

```
bonsai:prov/dataset_0 a prov:Entity ;
        dc:title "Exiobase"
        rdfs:label "LCSA dataset by the EXIOBASE-Consortium, version 3_3_17";
        dc:date "2019-03-12"^^xsd:date;
        dc:license <https://www.exiobase.eu/index.php/terms-of-use> ;
        dc:rights "Copyright©2015 - EXIOBASE Consortium" ;
        owl:versionInfo "3.3.17" ;
        prov:wasAttributedTo bonsai:foaf/provider_0.
```

Fig. 6. Provenance record of the Exiobase dataset v3.3.17. The PROV-O *Entity* records the specific version of Exiobase (i.e., v3.3.17), and the attribution to the EXIOBASE Consortium using the W3 PROV-O predicate `prov:wasAttributedTo`.

6 BONSAI Database In-Use

Currently, we have published linked open data obtained from the integration of the Exiobase and YSTAFDB datasets. This includes 15.3 M 49 K flows, 164 and 9 flow objects, 49 and 1686 activities, and 200 and 525 locations for Exiobase and YSTAFDB, respectively. In the following, we shortly describe how the BONSAI database can be used in practice for the calculation of environmental emissions. We also describe how the availability of provenance annotations allows us to inspect and assess the reliability of the provided information.

Example Use Case. A typical use of LCSA is the calculation of environmental emissions for industries and products of interest. For example, to estimate the environmental emissions related to production and consumption of steel products in China. In the following, we show which information we have access to by querying the BONSAI database. The queries are executed on the BONSAI SPARQL endpoint[3]. From Exiobase we obtain that ∼1237 megatonnes of steel were produced in China in the year 2011. Then we inspect how that steel was used and by whom: China uses approximately 92.7% of its domestic steel production. The database currently records that out of the whole of China's steel production ∼617 megatonnes (∼50%) were consumed by manufacturers of steel products, ∼127 megatonnes (∼10% of production) were consumed by manufacturers of electrical machinery, and ∼168 megatonnes (∼14% of production) were consumed by manufacturers of fabricated metal products. The database also contains information about environmental emissions, such as carbon dioxide, particulate matter (PM 2.5, PM 10), and other such emissions. The database shows that the Chinese steel production contributes to ∼1617 megatonnes of carbon dioxide and ∼1.77 megatonnes of particulate matter. Similar to the above

[3] Available at http://odas.aau.dk.

example, data on many other products of major industrial and agricultural sector can be extracted from Exiobase for the year 2011. Instead, YSTAFDB provides data on a smaller set of 62 elements and various engineering materials, e.g., steel, but on more granular spatial scales and timeframes, ranging from cities to global and from the 1800s to 2013. Therefore, the combined information from the two datasets can be used to make more qualified environmental decisions regarding the environmental performance of steel production, such as comparing the emissions from Chinese steel production to the national environmental emission quotas. Similarly, it can compare the impacts of Chinese steel production to other countries.

Provenance. In the above example, we found that China produced ∼1926 tonnes of steel in 2011. One typical question is to verify whether the most current version of the data is available as well as what is the source of such a datapoint. To address this, we query the provenance of our data to find the origin of the information on which we calculated the Chinese *steel* production. The first query finds the named graphs with *Flows* used in the calculation of the Chinese production of *steel*. The query is illustrated in Fig. 7. It finds all *Flows* which are the output of an *Activity*, where the *Activity* has an *Activity Type* labeled as *Manufcature of basic iron and steel*, and the *Activity* was performed in *China* (*Location*). To identify the origin of this information, it finds all named graphs to which such *Flows* belong. In our example, we find that all extracted *Flows* of the product *Iron* used in the calculation of the Chinese production origin from the named graph `bonsai:data/exiobase_3_3_17/hsup`. The SPARQL DESCRIBE query can now be used to describe the resources, which for this example results in the record illustrated in Fig. 8.

```
SELECT DISTINCT ?collection
FROM ...
WHERE {
  ?flows a bont:Flow .
  ?flows bont:outputOf ?activity .
  ?activity bont:activityType / rdfs:label "Manufacture of steel...".
  ?activity bont:location / rdfs:label "CN".
  ?collection prov:hadMember ?flows }
```

Fig. 7. Query fragment (reduced for space constraints) for finding the collections (e.g., named graphs) where flows regarding Chinese steel production origin from.

As illustrated in the figure, the provenance relation `prov:wasGeneratedBy` shows that the graph was generated by a data extraction activity identified by `bonsai:prov/dataExtractionActivity_0`. We further query the provenance of the database to investigate this data extraction activity (also in Fig. 5). The record has a provenance relation `prov:used` to the dataset located at `bonsai:prov/dataset_0`, which means that this dataset was used in the activity to create the named graph located at `bonsai:data/exiobase_3_3_17/hsup`.

```
bonsai:data/exiobase3_3_17/hsup a prov:Collection ;
    prov:generatedAtTime "2019-12-02"^^xsd:date ;
    prov:wasAttributedTo bonsai:foaf/bonsai ;
    prov:wasGeneratedBy bonsai:prov/dataExtractionActivity_0 ;
    prov:hadMember  bonsai:data/exiobase3_3_17/hsup/f_9109,
        bonsai:data/exiobase3_3_17/hsup/f_9096,
        ...
```

Fig. 8. Record for the collection `bonsai:data/exiobase_3_3_17/hsup`. It encodes provenance metadata for *Flows*, *Activities* and *Locations* extracted from the dataset `exiobase_3_3_17`. Non-provenance metadata has been omitted from the figure.

Hence, when provenance of the dataset is queried (illustrated in Fig. 6), we find the PROV-O relation `prov:wasAttributedTo`, which is an attribution from the dataset *Entity* to the *Agent* responsible for its delivery. Hence, we query the database again to find information about the *Agent* with URI `bonsai:foaf/provider_0`, as referred to in the relation. The resulting record is illustrated in Fig. 9. This allows us to reach directly the source of the dataset and its publisher to verify the currently available information. Moreover, the provenance information about the extraction activity also has a `prov:hadPlan` relation to an extraction script entity identified with `bonsai:prov/extractionScript` (see Fig. 5). This entity points to a specific version of a GitHub repository containing the version of code used for the extraction of *Flows*, *Activity Types*, *Flow Objects*, and *Locations*. Hence, provenance for the data extraction script is also maintained.

```
bonsai:foaf/provider_0 a prov:Agent, org:Organization ;
    dc:description "Consortium creating datasets for LCSA" ;
    dc:title "Exiobase Consortium" ;
    foaf:homepage <https://www.exiobase.eu/> .
```

Fig. 9. Provenance record for the PROV-O *Agent* `bonsai:foaf/provider_0`. The record contains information about the Exiobase Consortium.

Triplestore Performance. To publish the data we collected through an open SPARQL endpoint, we first deployed Jena as our triple store. Yet, during initial tests with different amounts of data, we quickly witnessed that uploading all our triples (∼15 M) proved unfeasible. In particular, as we tried to store more data, the disk space required by Jena was increasing faster than expected. Furthermore, queries to address the competency questions [6] were requiring many minutes to compute, even when processing only a subset of the data. Therefore, we investigated alternative options in terms of triplestore performance for our domain specific case of LCSA. In particular, we compared Jena and Open Virtuoso using the LITMUS benchmark framework [19]. The benchmark was run

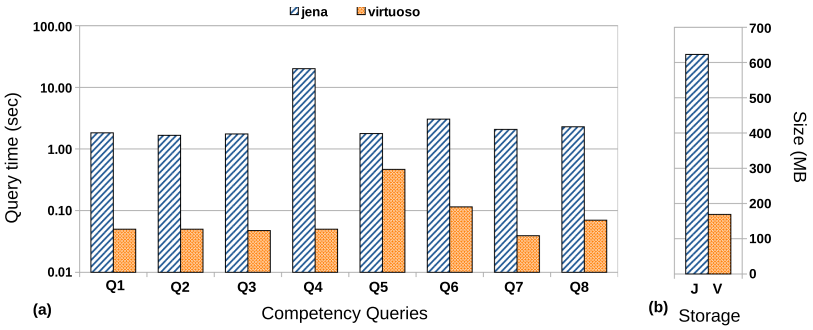

Fig. 10. Comparison of query time and disk footprint. The label under each column, corresponds to queries related to the competency questions [6].

on a virtual server with 8 cores and 64 GB of RAM. Our benchmark adopted the SPARQL queries converted from the competency questions over a subset of our full dataset. In our results (see Fig. 10), Open Virtuoso greatly outperformed Jena in all queries by an order of magnitude in running time. Moreover, Jena storage files on disk had 3× larger space footprint than Virtuoso. Therefore, we concluded that Open Virtuoso was the best choice for our needs and is now used as the DBMS for our SPARQL endpoint.

7 Lessons Learned and Future Work

In this work we have presented how we employed Linked Open Data and Semantic Web technologies for achieving the goal of integrating LCSA datasets. This allowed us to establish the first Linked Open Data database for product footprinting (See footnote 3). The current implementation overcomes several limitations in previous similar efforts. In the following, we present a brief summary of both the advantages and challenges that we encountered in this process.

Advantages of Semantic Web. We demonstrate the benefits of employing Semantic Web technologies to support open and transparent LCS Analysis. To achieve its full potential, LCSA requires the collaboration and sharing of information at many different levels both from governments and organizations. Their interoperability is of crucial importance for the effective computation of IO models. These models are required to investigate global and local impacts due to the change in demand for products and services. In particular, the adoption of a common ontology alongside established standards for data interoperability enables not only researchers and practitioners to have open access to environmental information, but also facilitates other providers to contribute to the database by sharing their own data.

Moreover, we prove the advantages of Semantic Web technologies in the domain of LCSA, by successfully integrating the two datasets: Exiobase and YSTAFDB. These datasets are now fully interoperable and can be queried

and analyzed altogether. Further, we plan to exploit novel SW technologies by extending the pipeline with a data consistency checking process using SHACL constraints. This represents a unique and unprecedented opportunity for the future of LCSA.

Challenges. As the project grows, we expect new challenges in ensuring the computational scalability of the extraction pipeline. Currently, without the manual process for correspondence tables, the pipeline takes 8 hours to run, using a virtual machine with 8 cores and 64 GB RAM. Even though the complexity of the pipeline only grows linearly with respect to the number of triples in the accumulated datasets, runtime could become an issue when more and larger datasets are integrated. We plan to cope with this problem using parallelization techniques since the main processes in the pipeline are highly parallelizable.

Choice of Triple Store. As described in the previous section, the choice of data-management system was crucial to allow the necessary scalability of our database. While we first deployed Jena as our triple store, influenced by its popularity along with its open-source license, this choice revealed to be unfeasible. Open Virtuoso instead revealed to be a more solid choice. This was an important practical lesson for us.

User Interface. The use of SW technologies allows open access to the data for both humans and machines, thanks among others to the adoption of the RDF standard and SPARQL query language. Yet, SPARQL is hard to use for non-expert users. To bridge this gap we have deployed a simplified user interface by adapting the yasGUI client (See footnote 3). Also, we enhanced the GUI with query templates for easy access to common LCSA SPARQL queries. This interface presents a list of query templates, among which are present the examples adopted in this work and the competency questions defined with the domain experts. Despite the simplicity of the current GUI, it enabled LCSA experts to query the data in new ways. This led to find a flaw in a fundamental data assumption they were relying on in their handling of the data. Hence, Semantic Web technologies exploited through our GUI, allowed us to open the data enough for this assumption to be tested false by the domain experts and enabled the experts to design corrections in the data processing step. In future, we plan to implement advanced GUIs and as well as a Python library to be used within a data-science notebook to further empower domain experts in their analysis.

Provenance. In this work, we describe the data integration workflow established for the conversion of new datasets into interoperable Linked Open Data. This data will be used to derive complex models, hence we also require to verify the source of the data and the algorithms used in the calculation of the models. Hence, we employed the PROV-O ontology to implement provenance modeling of the entities, activities, and agents involved in the construction and updates of the database. Enabling the tracking of provenance information was one of the most important goals of this work and a key enabler for transparent and reliable LCSA. Nonetheless, while the PROV-O model is easy on the surface, the flexibility of the model presented a non-obvious challenge when deciding how to

adapt its vocabulary to our domain. In particular, it is not straightforward to decide the most convenient level of granularity at which to record provenance information. Finally, it was challenging to determine whether a specific provenance model meets all the necessary requirements. To this end, we designed a set of basic provenance competency questions, which we plan to expand in the future. The implementation of a model of provenance was done through multiple iterations and cross-referenced with the competency questions. In its current implementation, we focused on adopting the viewpoint of the domain experts, who are used to handle datasets in terms of files and data-providers.

Acknowledgments. This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 838216, the Independent Research Fund Denmark under grant agreement no. DFF-8048-00051B, and strategic research funding from Aalborg University Tech Faculty: ODA - Open Data for sustainability Assessment.

References

1. Buneman, P., Khanna, S., Tan, W.-C.: Data provenance: some basic issues. In: International Conference on Foundations of Software Technology and Theoretical Computer Science, pp. 87–93 (2000)
2. Celino, I.: Human computation VGI provenance: semantic web-based representation and publishing. IEEE Trans. Geosc. Remote Sensing **51**(11), 5137–5144 (2013)
3. Curran, M.A.: Environmental life-cycle assessment. Int. J. Life Cycle Assess. **1**(3), 179–179 (1996)
4. De Nies, T., et al.: Git2PROV: exposing version control system content as W3C PROV. In: ISWC, pp. 125–128 (2013)
5. Galárraga, L., Mathiassen, K.A.M., Hose, K.: QBOAirbase: the European air quality database as an RDF cube. In: International Semantic Web Conference (Posters, Demos & Industry Tracks) (2017)
6. Ghose, A., Hose, K., Lissandrini, M., Weidema, B.P.: An open source dataset and ontology for product footprinting. In: ESWC, pp. 75–79 (2019)
7. Jansen, P.K., Raa, T.T.: The choice of model in the construction of input-output coefficients matrices. Int. Econ. Rev. **31**(1), 213–227 (1990)
8. Kuczenski, B., Davis, C.B., Rivela, B., Janowicz, K.: Semantic catalogs for life cycle assessment data. J. Clean. Prod. **137**, 1109–1117 (2016)
9. Lebo, T., et al.: PROV-O: The PROV ontology. W3C recommendation (2013)
10. Merciai, S., Schmidt, J.: Methodology for the construction of global multi-regional hybrid supply and use tables for the EXIOBASE v3 database. J. Ind. Ecol. **22**(3), 516–531 (2018)
11. Messerli, P., et al.: Global Sustainable Development Report 2019: The Future is Now-Science for Achieving Sustainable Development (2019)
12. Myers, R.J., Reck, B.K., Graedel, T.: YSTAFDB, a unified database of material stocks and flows for sustainability science. Sci. Data **6**(1), 1–13 (2019)
13. Nolin, M.-A., et al.: Bio2RDF network of linked data. In: Semantic Web Challenge; ISWC 2008 (2008)
14. Prud'hommeaux, E., Labra Gayo, J.E., Solbrig, H.: Shape expressions: an RDF validation and transformation language. In: Proceedings of the 10th International Conference on Semantic Systems, pp. 32–40 (2014)

15. Sala, S., Ciuffo, B., Nijkamp, P.: A systemic framework for sustainability assessment. Ecol. Econ. **119**, 314–325 (2015)
16. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM Sigmod Rec. **34**(3), 31–36 (2005)
17. Suh, S., Huppes, G.: Methods for life cycle inventory of a product. J. Clean. Prod. **13**(7), 687–697 (2005)
18. Suh, S., Weidema, B., Schmidt, J.H., Heijungs, R.: Generalized make and use framework for allocation in life cycle assessment. J. Ind. Ecol. **14**(2), 335–353 (2010)
19. Thakkar, H.: Towards an open extensible framework for empirical benchmarking of data management solutions: LITMUS. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10250, pp. 256–266. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58451-5_20
20. Visotsky, D., Patel, A., Summers, J.: Using design requirements for environmental assessment of products: a historical based method. Proc. CIRP **61**, 69–74 (2017)
21. Weidema, B.P., Schmidt, J., Fantke, P., Pauliuk, S.: On the boundary between economy and environment in life cycle assessment. Int. J. Life Cycle Assess. **23**(9), 1839–1846 (2018)
22. Yolanda Gil, S.M.: PROV Model Primer. https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/. Accessed 12 May 2019