An Open Source Dataset and Ontology for Product Footprinting

Agneta Ghose[1*], Katja Hose[2*], Matteo Lissandrini[2*✉], and Bo Pedersen Weidema[1*]

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

Abstract. Product footprint describes the environmental impacts of a product system. To identify such impact, Life Cycle Assessment (LCA) takes into account the entire lifespan and production chain, from material extraction to final disposal or recycling. This requires gathering data from a variety of heterogeneous sources, but current access to those is limited and often expensive. The BONSAI project, instead, aims to build a shared resource where the community can contribute to data generation, validation, and management decisions. In particular, its first goal is to produce an open dataset and an open source toolchain capable of supporting LCA calculations. This will allow the science of lifecycle assessment to perform in a more transparent and more reproducible way, and will foster data integration and sharing. Linked Open Data and semantic technologies are a natural choice for achieving this goal. In this work, we present the first results of this effort[3]: (1) the core of a comprehensive ontology for industrial ecology and associated relevant data; and (2) the first steps towards an RDF dataset and associated tools to incorporate several large LCA data sources.

1 Introduction

Life Cycle Assessment (LCA), also called "product footprinting", is concerned with analyzing the environmental impact of products, taking into account their complete production chain and lifespan [1]. For instance, assessing the impacts of operating a solar array goes beyond the pure manufacturing and assembly of the photo-voltaic modules. It also includes all impacts and emissions relative to the extraction of raw materials, transportation, installation, operation, and the final disposal. Hence, to produce an LCA in this case, it first requires the gathering of all relevant data from different sources into a so-called Life Cycle Inventory (LCI). Then, such data can be integrated and processed with state-of-the-art models and procedures. LCA is a highly complex and interdisciplinary field that requires synthesizing information from a variety of discipline-specific studies. Nonetheless, it has a fundamental role in the realization of a sustainable world where human needs are met while minimizing the harm to the environment and without reducing the ability of future generations to meet their needs [4].

* Authors are listed in alphabetical order [3] https://github.com/BONSAMURAIS/

To a large extent, LCA currently exploits large background databases, often proprietary, which are expensive to access and consequently provide limited access to both the data and decisions on its management. Therefore, given the transversal importance of LCA, following the principles of Open and FAIR data [5], there is the requirement to establish an Open Source dataset for product footprinting. While past studies have outlined compact ontologies that formalize the spatio-temporal scope of activities in LCA [3, 6], those are limited in their modeling of the domain [4] and have not resulted in the publication of open datasets. This work, led by the BONSAI (`https://bonsai.uno`) non-for-profit association, plans to overcome the limitation of previous initiatives. Here, we describe the first results of this effort, which involves experts and companies in the sector of environmental assessment and sustainability planning, and the long-term plan for the first open dataset and ontology for product footprinting.

2 Product Footprints: Development, Ontology, and Data

The BONSAI initiative has three main objectives: (1) the definition of a comprehensive ontology for industrial ecology (IE), (2) the publication and maintenance of an open source IE dataset for LCA, and (3) the development of a toolbox for data ingestion, integration, validation, and sharing to maintain such a dataset.

2.1 The BONSAI Ontology and Data

Domain and Purpose: The BONSAI dataset and its accompanying ontology describe entities that play important roles in representing the environmental impacts associated with all the stages of a product's life. To identify the entities and concepts that are expected, we defined a number of competency questions. Example competency questions include: (i) Is the flow x a determining flow for activity y (e.g., electricity from a power plant)? (ii) What is the amount of flow x emitted as output during the time period y (e.g., the emission of landfill gas)? (iii) What is the location of the agent performing activity y (e.g., where is the coal power plant located) and what other agents performing the same type of activity are present in the same location?

The competency questions highlight the centrality of two concepts: the Flow (e.g., some steel being produced, some coal being consumed) and the associated Activity (the production of steel). In LCA models, each activity is a *consequence* of a specific determining flow (e.g., the activity of steel production is a consequence of the demand for the flow of steel), while some other flows are subordinate (e.g., the consumption of coal and the CO_2 emission).

Ontology building: To build the domain ontology we interacted with experts and analyzed existing datasets. We started from EXIOBASE (`www.exiobase.eu`), a well established database (with a tabular model) comprising, among others, 43 countries, 200 products, and 163 industries. Data published by international initiatives – Food and Agriculture Organization (FAO), United Nations Environment Programme (UNEP) – is also frequently used by LCA practitioners. We expanded existing ontologies drafted in the same context [3, 6]. These original proposals did not provide (a) an adequate vocabulary for expressing all

the required details of a flow and (b) clear linking to other relevant ontologies. We also identified a set of relevant ontologies and databases in complementary domains (e.g., units of measure[1], time[2], and GeoNames). In particular, we *identified important terms* like Flow, Activity, Input/Output, and Agent and we *defined classes and class hierarchies* with the most important terms being top-level classes (see Figure 1). We note that we have explicitly established mappings with other interconnected ontologies and vocabularies (e.g., Schema.org) in order to foster data integration, discovery, and alignment. In the future, we plan to expand and integrate other vocabularies, e.g, vocabularies for data provenance[3] and statistical data[4]. We have *defined properties for each class*, specifying the instances of classes representing allowed domain and range values. Among others, an important aspect is that each Activity must be associated with at a least one Flow that is classified as Determining Flow.

Data Extraction, Ontology Evaluation, and Documentation: The appropriateness of the ontology has been tested by a technical evaluation where 12 domain experts assessed the correctness and expressive power of the available definition against the reference dataset. A conversion tool has been developed to process the EXIOBASE data and produce the corresponding RDF data. The set of competency questions were revised and used, with corresponding SPARQL queries, to verify the appropriateness of the model. The ontology and the dataset are accompanied with an external "living" documentation that describes among others, the ontology purpose, class definitions, description of class properties, and evaluation (at `https://github.com/BONSAMURAIS`).

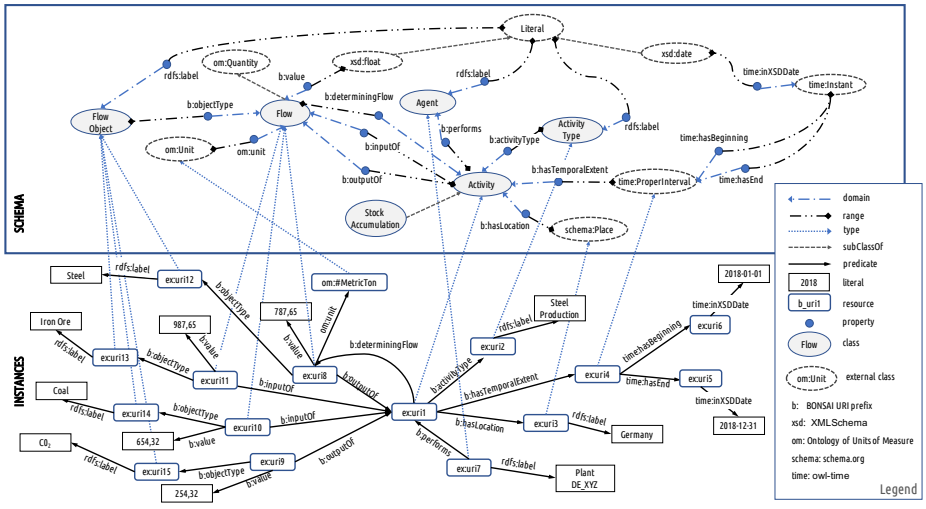

Fig. 1. The core of the ontology and an example instantiation.

Overview of the ontology: The main classes Activity and Flow have broad

[1] `www.ontology-of-units-of-measure.org` [2] `www.w3.org/TR/owl-time/`
[3] `www.w3.org/TR/prov-o/` [4] `www.w3.org/TR/vocab-data-cube/`

definitions according to the literature [3, 4], which facilitate their use with data supplied from external sources. Core concepts are defined as follows (Figure 1).

Activity is the act of doing within a temporal interval, this includes both human activities (e.g., production, consumption, and market activities) and environmental mechanisms (e.g., radiative forcing, pollination). *Agent* is defined as an entity (person or thing) that performs an activity. An agent has a location and the location of the activity is also determined by the agent performing it. *Flow* is defined as an entity that is produced or consumed by activities or stored within an activity (e.g., stock). *Determining Flow* is the flow of an activity determining its primary function. All other flows are co-produced by or demanded for in that specific activity but do not determine its existence. Usually, a change in the determining flow will affect the volume of all other flows involved.

2.2 Tools and Future Work

The long term plan is to allow the dataset to evolve and to third parties to contribute to it. Therefore, the project will provide tools building upon the state of the art [2] (i) to extract data from published studies and databases, (ii) to normalize to common industry and product classifiers, (iii) to assess data quality for many types of industrial ecology facts, and (iv) to build interpolation models for data across time and space.

3 Conclusions

Effective sustainability assessment requires access to data from a variety of heterogeneous sources. We believe that this effort will ensure low barriers for contributions from non-experts and for cross-dataset editing and that it will greatly benefit from the expertise and capabilities of the semantic web community.

Acknowledgments. This research was partially funded by the Danish Council for Independent Research (DFF) under grant agreement no. DFF-4093-00301B and Aalborg University's Talent Programme.

References

1. M. A. Curran. Environmental life-cycle assessment. *The International Journal of Life Cycle Assessment*, 1(3):179–179, 1996.
2. A. Harth, K. Hose, and R. Schenkel, editors. *Linked Data Management*. Chapman and Hall/CRC, 2014.
3. K. Janowicz, A. A. Krisnadhi, Y. Hu, S. Suh, P. Weidema, B. Rivela, J. Tiv, D. E. Meyer, P. Hitzler, W. Ingwersen, et al. A Minimal Ontology Pattern for Life Cycle Assessment Data. In *WOP*. CEUR-WS.org, 2015.
4. B. P. Weidema, J. Schmidt, P. Fantke, and S. Pauliuk. On the boundary between economy and environment in life cycle assessment. *The International Journal of Life Cycle Assessment*, 23(9):1839–1846, Sep 2018.
5. M. D. Wilkinson, M. Dumontier, and et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.
6. B. Yan, Y. Hu, B. Kuczenski, K. Janowicz, A. Ballatore, A. A. Krisnadhi, Y. Ju, P. Hitzler, S. Suh, and W. Ingwersen. An ontology for specifying spatiotemporal scopes in life cycle assessment. In *Diversity++@ ISWC*, pages 25–30, 2015.