The Sustainability Data Science Life Cycle for automating multi-purpose LCA workflows for the analysis of large product portfolios

Daniel Wehner[1*], *Tobias* Prenzel[1,2], *Thomas* Betten[2], *Ann-Kathrin* Briem[2], *Sun Hea* Hong[2] and *Robert* Ilg[1]

[1]Fraunhofer Institute for Building Physics IBP, Department Life Cycle Engineering, 70563 Stuttgart, Germany
[2]University of Stuttgart, Institute for Acoustics and Building Physics (IABP), 70563 Stuttgart, Germany

Abstract. Life Cycle Assessment (LCA) is a powerful and sophisticated tool to gain deep understanding of the environmental hotspots and optimization potentials of products. Yet, its cost-intensive manual data engineering and analysis workflows restrain its wider application in eco-design, green procurement, supply chain management, sustainable investment or other relevant business processes. Especially for large product portfolios and increasing reporting requirements, traditional LCA workflows and tools often fail to provide the necessary scalability. The Sustainability Data Science Life Cycle (S-DSLC) is a concept for workflow automation for multi-purpose LCA of large product portfolios. The concept integrates the frameworks of LCA, the cross-industry standard process for data mining (CRISP-DM), and the Data Science Life Cycle (DSLC). Key aspects of the concept are deep business-, stakeholder and user-understanding, deployment of LCA results in interactive browser tools (i.e. LCA-dashboards and Guided Analytics) tailored to the needs of individual roles and business processes, as well as the automation of data preparation, model generation and Life Cycle Impact Assessment based on modern data analytic tools. The demonstration of the concept shows substantial scalability improvements for dealing with large product portfolios and broad application of LCA results in various business processes.

1 Introduction

Sustainability management in business has to deal with an increasing amount of regulations and stakeholder requirements that call for scalable sustainability analysis capacity. On the one hand, enterprise internal stakeholders need particular product related environmental information to drive sustainable change by informed decisions in product development, procurement or strategy. On the other hand, external stakeholders (in particular clients and investors) request environmental product information needed to comply with an increasing number of requirements and regulations, such as the Corporate Sustainability Reporting

* Corresponding author: daniel.wehner@ibp.fraunhofer.de

Directive (CSRD) [1] or the Sustainable Finance Disclosure Regulation (SFDR) [2]. This environmental product information may have to be provided in compliance with the standards for product environmental footprints (PEF) [4], environmental product declarations (EPD) [4] or other standards based on Life Cycle Assessment (LCA) according to ISO 14040 [5].

While LCA [5] provides an appropriate approach to assess the environmental impacts of single products, it shows poor scalability when applied to assessing large product portfolios product-by-product. In particular, it neglects that the information required for an LCA study may not be available and fails to provide efficient solutions for deriving this information by blending and transforming data that are potentially only available in inconsistent formats or structures. Besides poor efficiency for dealing with a large number of products, another scalability bottleneck lies in the deployment of results by various business users throughout an enterprise, where traditional LCA reports fail to provide appropriate decision support for individual business processes [6].

In contrast, data science approaches such as the cross-industry standard for data mining (CRISP-DM) [7] or the data science life cycle (DSLC) [8] place a particular focus on practical solutions for typical data analysis challenges as well as the appropriate deployment of data analysis results for business users. While the general usefulness of data science approaches in the context of sustainability analysis has already been demonstrated [9][10][6], there is a lack of practical guidance for their implementation in the context of LCA-based multi-purpose analysis of large product portfolios. To the knowledge of the authors, no such approach has been presented to date.

2 The Sustainability Data Science Life Cycle (S-DSLC)

Addressing this lack, an integrated approach is proposed: the Sustainability Data Science Life Cycle depicts a standard approach for scalable sustainability analytics and product stewardship in a corporate environment (Fig. 1).

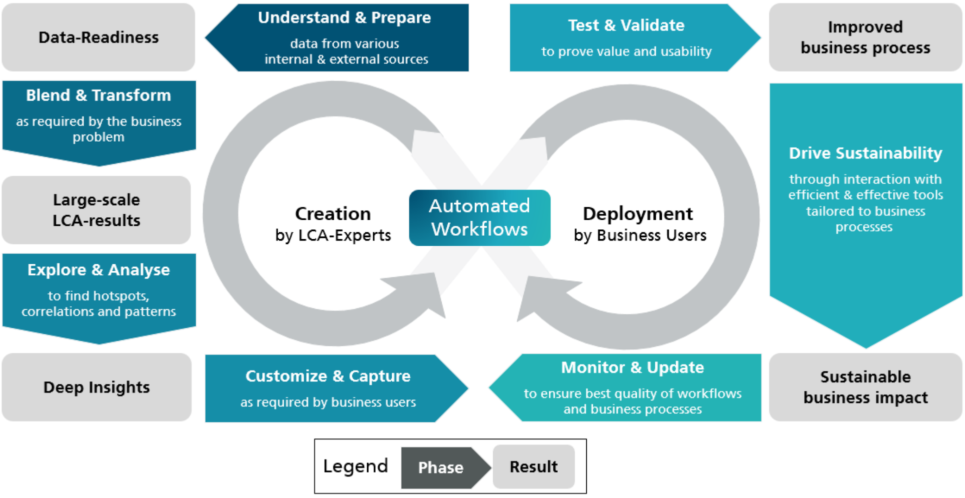
Data-Readiness
Understand & Prepare
data from various internal & external sources
Test & Validate
to prove value and usability
Improved business process
Blend & Transform
as required by the business problem
Large-scale LCA-results
Explore & Analyse
to find hotspots, correlations and patterns
Deep Insights
Creation
by LCA-Experts
Automated Workflows
Deployment
by Business Users
Drive Sustainability
through interaction with efficient & effective tools tailored to business processes
Customize & Capture
as required by business users
Monitor & Update
to ensure best quality of workflows and business processes
Sustainable business impact
Legend
Phase
Result

Fig. 1. Process model of the Sustainability Data Science Life Cycle (S-DSLC) describing the creation of automated workflows for deployment business users to drive sustainability.

The S-DSLC integrates the LCA framework (ISO 14040) with the two data science standards CRISP-DM [7] and the DSLC [8]. It consists of seven phases describing the creation of automated workflows by LCA experts that capture deep sustainability insights for deployment by business users without LCA expertise. Each phase or workflow part (arrows) results in a different level of insight (boxes) being crucial for the integral implementation of the S-DSLC and builds on the original phases of the three underlying standards (LCA, CRISP-DM & the DSLC).

The goal of the first phase (**understand & prepare**) is to achieve a general data readiness with regard to different product and production related sustainability questions throughout the enterprise. It comprises the understanding of sustainability related business problems and the data available to solve them. In line with the basic data requirements according to ISO14040, this typically involves at least primary data on products and their production as well as secondary LCA data (sometimes also referred to as background data) [5]. Various other data could be included if required by the scope of the S-DSLC implementation. Moreover, this phase also comprises the technical implementation of workflows to access the available data, to assess data quality and to automate data cleaning and preparation tasks that are of general relevance for multiple applications of data. Typical examples in the S-DSLC context are unit and data type conversions, harmonization of expressions or the extraction of required information from product identification codes or natural language descriptions.

In the second phase (**blend & transform**), the different data sources are blended and transformed to compute LCA results on a large-scale, e.g., for every single item of a large product portfolio. This comprises at least the implementation of mapping algorithms to link primary data on products and production with secondary LCA data as well as the required methods to compute LCA results. Depending on the complexity and the targeted level of detail, this may include solving linear systems of equations to scale the processes of an integrated production network or the algorithms for the life cycle impact assessment of life cycle inventories.

In the third phase (**explore & analyse**), LCA-experts analyse and explore the large-scale LCA results in order to extract sustainability insights as well as to identify data quality issues that only become obvious after data blending, in particular mapping errors and data gaps. Typically, advanced data analysis techniques are employed to deal with the large result space. This may include interactive dashboard tools for visual analysis, statistical methods or algorithms to identify patterns and correlations. It may also involve further transformations of the resulting data (i.e. the large-scale LCA results) as required to answer particular questions.

The target of the fourth phase (**customize & capture**) is to capture particular sustainability insights customized for the optimization of individual business processes to drive sustainable change, i.e. to create highly automated workflows business users can deploy with low effort and without LCA expertise. The implementation of such workflows may comprise dedicated dashboard tools, simple multi-step web-browser wizards, Application Programming Interfaces (APIs), or other applications and services. The customization step may require the implementation of further or re-designed data preparation, blending, transformation steps, the development of further models to enable a particular business process, or the optimization of the robustness of the automated workflow for dealing with new data.

Phase five (**test & validate**) is to ensure the necessary quality of the automated workflows before they are independently employed by business users to drive sustainable decision-making without the active involvement of LCA experts. Besides the general usefulness of the automated workflows, the LCA experts should also closely monitor the robustness of the results for new data and the correct interpretability by business users.

In phase six (**drive sustainability**) the approved automated workflows are made available for general and autonomous use in the targeted business processes to drive sustainable change by informed decisions.

Phase seven (**monitor & update**) is to ensure high quality of the deployed automated workflows. In the S-DSLC context, in particular the mapping algorithms of product, production and secondary LCA data may require regular updates, e.g., when new production processes or supplied goods have to be integrated or new background database versions are released. As neglecting such updates may lead to significant quality issues, LCA experts should monitor and manage automated workflows with respect to typical data quality indicators in the context of LCA, e.g., the ILCD data quality indicators [11].

3 Implementation example for green supply chain development

A solution for a fictive yet realistic business problem is implemented in the KNIME Analytics platform [12] to demonstrate the application of S-DSLC. The implementation use case is based on randomly generated data in structures typically found in business. For the use case, a fictive procurement department of a machine producing company wants to improve the sustainability of its sourcing processes for semi-finished steel products and thus to identify their suppliers providing the steel goods with the largest lever for improvements in the sourcing portfolio. As the future development of the company's own product sales heavily influences sourcing decisions, the analysis should be able to reflect respective sales trends.

To create an automated workflow that supports the decision-making for the described business problems various data are needed: (1) Data on current sales and trends, (2) Bill-of-Materials and related manufacturing data providing information on which products covered in the sales data require which semi-finished steel product, (3) sourcing data providing information on the suppliers of the respective semi-finished steel products, and (4) LCA data to estimate the environmental impacts of the sourced goods. For the implementation use case, data for (1), (2) and (3) were randomly generated in structures typically found in industrial enterprise resource planning (ERP) systems. The data cover fictive information on 28 machines totally consisting of roughly 500,000 parts sourced by 300 suppliers in 80 locations. Data are stored in a SQL database representing an ERP system. Fraunhofer IBP internal data were used to represent the required LCA data (4). Figure 2 depicts the automated workflow developed to support the business problem.

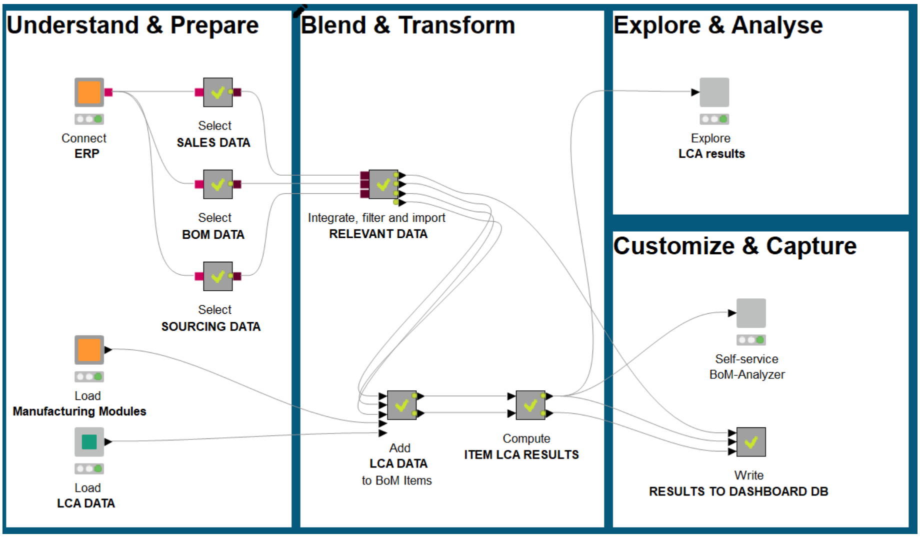

Fig. 2. Automated workflow implementing the S-DSLC phases "understand & prepare", "blend & transform", "explore and analyse" as well as "customize & capture" for a fictive use case in green supply chain development (Screenshot from KNIME Analytics Platform [12]).

The automated workflow pulls the required data from the ERP system and performs basic preparation and blending tasks as in database operations. Moreover, it loads LCA-data as well as additional data modelling in-house manufacturing processes (understand & prepare). All loaded data are blended based on a mapping algorithm, so that each item is matched with related primary or secondary LCA data, enabling the computation of item level LCA results (blend & transform). The results are forwarded to a workflow component providing explorative analysis functions, supporting LCA experts in the interpretation of the entire result space (explore & analyse). Furthermore, the workflow transforms and stores relevant data for improving sustainable sourcing decisions in a dedicated database feeding a respective business user dashboard (customize & capture). Figure 3 depicts an exemplary dashboard tool based on these data and ready for testing and validation in context of the targeted business process.

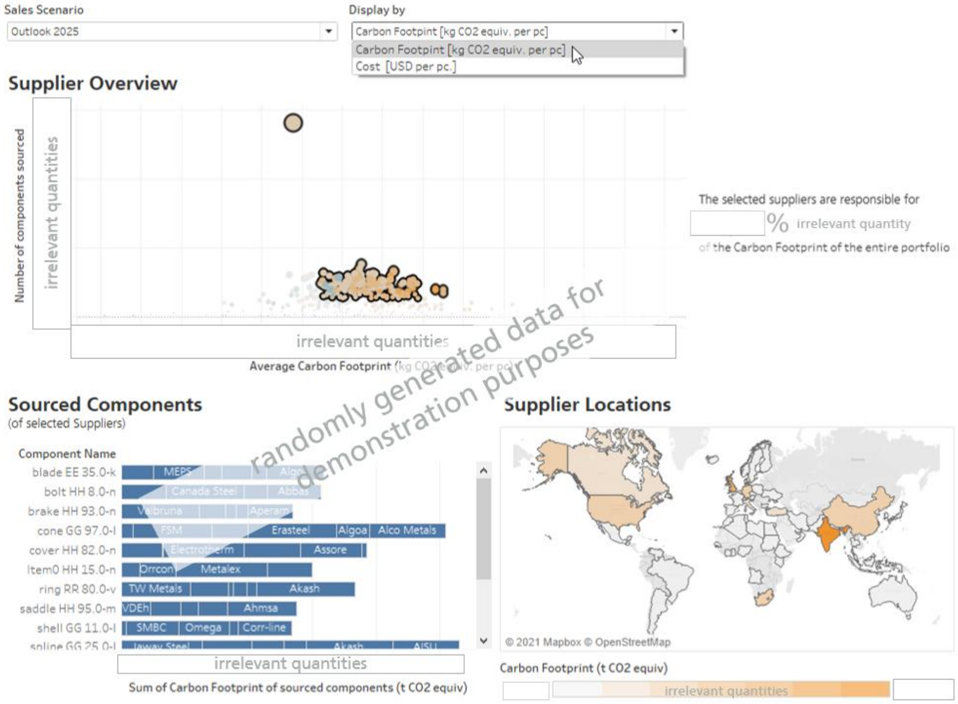

Fig. 3. Sample procurement dashboard facilitating green supply chain development, based on randomly generated data for demonstration purposes (therefore the resulting quantities are irrelevant).

The dashboard ranks suppliers by sustainability lever on product portfolio level, enabling users to select suppliers by the number of components and their average carbon footprint per piece (top left diagram). The share of the carbon footprint of the components sourced from the selected supplier(s) compared to the total carbon footprint of all sourced components is computed based on the selection and displayed in the top right corner. The bottom part of the dashboard gives further information on the components with the highest contribution to the portfolio carbon footprint, from which suppliers they were sourced and where the suppliers are located.

4 Conclusion and Outlook

The S-DSLC approach and its exemplary implementation in the context of green supply chain development show how core methodologies from LCA and data science can be used to substantially improve the scalability of LCA in business. The workflow of the implementation use case demonstrates how all data engineering processes from data access to the computation of item level LCA results can be automated, hence, eliminating a major scalability bottleneck for the LCA of large product portfolios. Moreover, it demonstrates how data science tools can be used to create easy-to-use applications for business users, hence, to scale-up the use of LCA throughout an enterprise. Consisting of several reusable building blocks for data access, basic cleaning operations or the computation of LCA results, the demonstrated workflow can be re-executed (e.g. to automatically compute updated results based on new data entries in the ERP system) or adapted for other applications (e.g., in product development or sustainability reporting). Furthermore, the visual modelling of the automated workflow offers good maintainability, for example, when mapping algorithms need to be updated or new data sources are to be integrated. As the implementation use case did not demonstrate the S-DSLC phases “explore & analyse”, “test and validate” as well as “monitor and update”, this should be object of further investigation.

References

1. European Commission, *Proposal for a directive of the European Parliament and of the Council amending Directive 2013/34/EU, Directive 2004/109/EC, Directive 2006/43/EC and Regulation (EU) No 537/2014, as regards corporate sustainability reporting*. (2021)
2. European Commission, *Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector*. (2019)
3. European Commission, *The development of the PEF and OEF methods*. Accessed online: https://ec.europa.eu/environment/eussd/smgp/dev_methods.htm (2021)
4. ISO 14025:2006, *Environmental labels and declarations, type III environmental declarations, principles and procedures*. (2020)
5. ISO 14040:2006, *Environmental management - Life cycle assessment - Principles and framework*. (2016)
6. D. Wehner, et al., *Ordnungsrahmen und Ansätze für das ökologische Risikomanagement bei der Produktentwicklung*. Forschungsergebnisse aus der Bauphysik, Band 40, Fraunhofer Verlag (2020)
7. P. Chapman, et al., *The CRISP-DM 1.0 Step-by-step data mining guide*. (1999)
8. M. Berthold, et al., *The Data Science Life Cycle: A New Standard for Operationalizing Data Science*. Accessed online: https://www.knime.com/blog/the-data-science-life-cycle-a-new-standard (2021)
9. T. Betten, et al., *Integration of Big Data Analytics into Life Cycle Assessment*, American Center for Life Cycle Assessment, Fort Collins, CO (2018)
10. D. Wehner, et al., *Towards industry 4.0 ready advanced sustainability analytics*, Challenges for Technology Innovation: An Agenda for the Future. CRC Press (2017)
11. European Commission, *ILCD Handbook*, Luxembourg (2010)
12. KNIME AG, *KNIME Analytics Platform – Creating Data Science*, Accessed online: https://www.knime.com/knime-analytics-platform (2021)