Athena Sustainable Materials Institute

Data Harmonization in the Impact Estimator Software

Final report

FII PSA S21-088-WF

March 31, 2021

This project is addressing harmonization and quality for the material data behind the free whole-building LCA software tool, the Athena Impact Estimator for Buildings. We describe the work to date and next steps in this brief report.

Objective

The objective of this project is data harmonization in the Impact Estimator, to ensure a level playing field for all materials in our tool and within Canada's national effort to create databases, methods and other resources that support wider and better use of LCA (the LCA^2 initiative). An additional objective is to introduce data quality indicators for material data used by the Impact Estimator.

Rationale and impacts

The Impact Estimator for Buildings is a software tool for life cycle assessment (LCA) of building projects. The Impact Estimator is provided for free as a public service by the non-profit Athena Institute. The tool simplifies life cycle assessment so that non-LCA experts such as architects and engineers can easily use it to help guide their design decisions. The Impact Estimator is both the only free tool and the only tool with regionally appropriate data for Canada, and thus it is Canada's de facto "national" whole-building LCA tool. It is also regionalized for the U.S. and is widely used there as well.

Reliability of LCA results is a significant concern for whole-building LCA in general. One big factor in reliability is consistency and quality of the background data and methods, an issue with all whole-building LCA software tools. While the Impact Estimator has a good track record for data consistency and quality, harmonization of our data is increasingly becoming a problem due to a proliferation of environmental product declarations (EPDs) with disparate results (sometimes EPDs are the only publicly available source of LCA data on products).

This harmonization and data quality effort will provide users with more information to help them gauge reliability of results, and it will assist in leveling the playing field for material comparisons in the Impact Estimator. This work will be particularly helpful for fair treatment of wood products, because the most recent LCA/EPD update of North American wood products creates an unfair disadvantage for wood compared to other products whose EPDs were done to a previous generation of background data and methodology (a change in the background has caused one of the LCA metrics for wood to be calculated much higher than before).

This work will significantly advance the reliability of whole-building LCA results. This will in turn improve the ability of the Impact Estimator to support policy, such as the embodied carbon provisions in the City of Vancouver rezoning policy. In addition, reliability of results will encourage wider use of LCA in design practice.

This work will be aligned with the Canadian national initiative LCA[2], which is developing a life cycle inventory database and other LCA resources to help Canada meet its climate change objectives.

The improved Impact Estimator will be freely available to the public. Data harmonization and data quality transparency will be beneficial to third-party software efforts as well. We are currently developing the capacity to provide LCA results as a back engine to software tools developed by others. This will help expand the accessibility of LCA for users.

Method
Develop a protocol for data harmonization and data quality indicators and begin to implement it. This work is in conformance with relevant ISO standards and in alignment with existing data quality assessment protocols.

Work completed to date
The work reported here is the first half of an effort that will continue for the next few months, with funding from another source.

We report work to date on these three tasks:

1. Data quality assessment
2. Data harmonization protocol
3. Documentation

<u>Task 1: Perform data quality assessment and apply indicators to a sample of our data (wood and cement).</u>

We identified a data quality assessment protocol and began applying qualitative indicators to the Athena dataset of over 200 materials. Data quality assessment aims to provide information to the user on the state of a dataset according to several indicators. At present, we are using the same indicators as used in the ecoinvent[1] database, while also allowing for other indicators to be added in the future, if need be. Scores range from 1 (most desirable) to 5 (least desirable), for the following five indicators: reliability, completeness, temporal correlation, geographical correlation, and further technological correlation. We explain these further below.

[1] Ecoinvent is a life cycle inventory database widely used in life cycle assessment (including by us in our product LCA work), and it will be the basis for the upcoming Canadian national life cycle inventory database. For the ecoinvent data quality methodology, see: Weidema B P, Bauer C, Hischier R, Mutel C, Nemecek T, Reinhard J, Vadenbo C O, Wernet G. (2013). Overview and methodology. Data quality guideline for the ecoinvent database version 3. Ecoinvent Report 1 (v3). St. Gallen: The ecoinvent Centre, Table 10.4

Reliability indicates the pedigree of how input data is measured when creating the dataset, ranging from a 1 for verified data based on measurements, to a 5 for non-qualified estimates. For example, our wood and cement datasets both rate 1 in this indicator because they are both based on rigorous surveys completed by the facility managers from their local material and purchasing records, and mass balanced and cross checked by the LCA practitioners.

Completeness indicates how representative the data is with regard to the overall market that the dataset represents, from a 1 where all manufacturing sites are surveyed over an adequate study period (usually 12 months of normal operations), to a 5 where the representativeness is unknown or from a small number of facilities and for shorter periods of operations. Our wood and cement datasets both rate a 2 because they are 12-month studies from facilities representing over 50% of their markets.

Temporal correlation indicates how old the survey data is, from a 1 if less than three years old to a 5 if more than 15 years or unknown. The wood LCA data was updated in 2018 but is based on primary data collected in 2015; because the primary data is more than six years old, this rates a 3. Similarly, for cement, the LCA was conducted in 2016 but is based on 2014 facility data, so it rates a 3.

Geographical correlation indicates if the data collection is from the same location that the dataset represents, from a 1 where the data is from the area under study to a 5 if the geographic area of the data is unknown or from a different area altogether. The wood and cement datasets both score a 1 because both LCAs were for Canadian wood and cement production.

Further technological correlation indicates how relevant the technology used in the facilities is compared to the technology that the dataset represents, from a 1 for data from enterprises, processes and materials under the study, to a 5 for data based on related processes in a laboratory or from different technology. Both the wood and cement datasets score a 1, almost by default, because they both are studies of current production facilities in their industries.

In summary, see Table 1 for the pedigree matrix from the ecoinvent data quality guidelines.

Table 1 - Data quality assessment indicators

Indicator score	**1**	**2**	**3**	**4**	**5** (default)
Reliability	Verified data based on measurements	Verified data partly based on assumptions ***or*** non-verified data based on measurements	Non-verified data partly based on qualified estimates	Qualified estimate (e.g. by industrial ex- pert)	Non-qualified estimate
Completeness	Representative data from all sites relevant for the market considered, over an adequate period to even out normal fluctuations	Representative data from >50% of the sites relevant for the market considered, over an adequate period to even out normal fluctuations	Representative data from only some sites (<<50%) relevant for the market considered ***or*** >50% of sites but from shorter periods	Representative data from only one site relevant for the market considered ***or*** some sites but from shorter periods	Representativeness unknown or data from a small number of sites ***and*** from shorter periods
Temporal correlation	Less than 3 years of difference to the time period of the dataset	Less than 6 years of difference to the time period of the dataset	Less than 10 years of difference to the time period of the dataset	Less than 15 years of difference to the time period of the dataset	Age of data unknown or more than 15 years of difference to the time period of the dataset
Geographical correlation	Data from area under study	Average data from larger area in which the area under study is included	Data from area with similar production conditions	Data from area with slightly similar production conditions	Data from unknown ***or*** distinctly different area
Further technological correlation	Data from enterprises, processes and materials under study	Data from processes and materials under study (i.e. identical technology) but from different enterprises	Data from processes and materials under study but from different technology	Data on related processes or materials	Data on related pro- cesses on laboratory scale ***or*** from different technology

Task 2: Develop a harmonization protocol and pilot test

There are a number of technical factors in the background of the LCA data for materials that can cause inconsistency between data sets. Inconsistent data on materials will lead to uncertainty in whole-building LCA results. In this task, we developed a draft protocol to harmonize our data.

In our view, the only way to do this properly is to go back to the fundamental life cycle inventory (LCI) data for each material and recalculate the LCA results consistently. This means we will model each dataset in Simapro (per functional unit of each material), all using the same background processes (e.g. electricity profiles), and calculate the life cycle impact assessment (LCIA) intensities in accordance with ISO 21930. These results will then be used to populate the background data in the Impact Estimator. Once all of our datasets are modelled in the same database, using the same background datasets, they can all be updated any time the background datasets are updated, and thereby always kept consistent.

We tested the approach on wood and cement. For these materials, because we did the original LCAs, we have access to the fundamental data (unit process data: flows for each individual material, energy etc.); these materials present no problem for our harmonization method. The method is straightforward to implement when we have access to LCI data (from our own LCAs or from others).

However, we will not always have such access. We often receive data from other sources, and the availability of the background data can be limited. Steel data is a good example of this; steel data is only available at the level of system processes (not the unit process data we need). System process data is a list of resources used and the emissions produced, but not which processes use those resources or produce those emissions.

For an explanation of why this is a problem, let's use electricity as an example. We know from the LCA that wood uses X kWh of electricity per m^3 of product; we can model that for our database, and also regionalize that data by modelling X kWhs in different regions of the country – we can also update the datasets easily when different regional electricity profiles are published. We cannot do this if all we have is system process data, as with steel. This data is static because the electricity profile is "baked in;" the data gives the results (MJ of coal, natural gas, hydro, kg of CO_2 etc) based on whatever electricity profile was used at the time of modelling. To get results for a different electricity profile, we need to know how many kWhs of electricity led to the system level results, and we cannot infer this from the available data. If we could, we would be able to convert system process data into unit process models. Instead, to harmonize materials with limited available data like steel, we would need to ask the providers of that dataset to rerun those models and give us new results whenever we are updating our other data, which isn't feasible. This means we will have to accept that some data in the Impact Estimator will not be consistent with the rest of our data, and we will have to flag this in some manner.

Task 3: Prepare draft documentation

We need to communicate to the public our method for data quality indicators and data harmonization, for transparency and traceability. We additionally will need other messaging to put the methodology in context for users.

For our data quality assessment indicators, the explanation in this report is a first draft of our documentation. We will consider how much further to develop the material, for example, perhaps to include justification of our ratings for specific materials. We additionally need to consider adding indicators to address quality and uncertainty concerns related to harmonization problems, and we will need to explain this.

For our data harmonization protocol, we will need more extensive documentation than captured briefly in this report. This is because we will be potentially modifying material data away from its published form, and that requires careful and transparent disclosure of method and rationale.

We identified additional user messaging to be developed regarding data quality and whole-building LCA. For example, there are limitations and uncertainty for any whole-building LCA results, particularly if there are data quality concerns. When we communicate data quality information to users, we would like to provide the educational context so that results from the Impact Estimator are used responsibly when there is a high degree of uncertainty. In other words, when the information we are conveying functions essentially as a red flag, we should ensure this message is clear and possibly provide guidance on how to act on the information.

We have begun to consider the mechanism for conveying data quality information to users. Our method will be captured in detail in our documentation for the Impact Estimator (user manual, web site, and embedded within the user experience in the new web version). Our intention is that users will be able to view data quality indicators for any material in our database. In addition, we hope to add data quality indicators to whole-building LCA results. For example, for temporal correlation, we might be able to add a label like this: "75% of your results come from datasets that are three years old or less, 14% that are six years old or less" etc.

Similarly, we intend to communicate an uncertainty message related to data harmonization. For example, we might put a label like this: "70% of your results come from materials that are the most recent version of our harmonized database, 25% from static system process data, and 5% from data provided by the user." Again, we will need to explain the relevance of this information.

Next steps:
We will continue assessing our database and applying data quality indicators per the method described here. We will also consider additional indicators as discussed in this report. We expect to have this task substantially completed by the time we next release the Impact Estimator; the next release will be the beta launch of our new web version (we are aiming for this summer), which is a significant overhaul of the software and will include many new features, including the work reported here. Our intention is that users will be able to view data quality indicators when using the Impact Estimator.

We will complete the harmonization of most materials and will include our updated database in the upcoming beta launch of the Impact Estimator. We need to consider how to indicate which material data is non-harmonized; we may decide to make this an additional data quality indicator. A subset topic here is data harmonization warnings if we allow users to include external material data (from Environmental Product Declarations) in their Impact Estimator results. Transparent communication on this issue might be a driver for the relevant industries to release the data we need for harmonization, and we will give this some thought as we develop our documentation and our communications plan.

A number of communication tasks will be addressed, including the specific nature of output to users (e.g., how we will include data quality information in the Impact Estimator reports that users generate) and the medium for communicating our protocols and guidance or educational context. Prior to the beta launch, we will develop beta documentation of our methods, recognizing that this will likely evolve as we get user feedback and refine our methods.

Report submitted by:
Jennifer O'Connor, President and Grant Finlayson, Senior Research Associate
Email: Jennifer.oconnor@athenasmi.org
Phone: 343-700-5250