GreenDelta
sustainability consulting + software

Data quality management in the MSDB database – what are good LCA datasets

openLCA conference, Berlin, April 22 2026

Dr. Andreas Ciroth
GreenDelta GmbH

Topics for the talk

- Reflecting reality in an LCA database
- The quality assessment in the MSDB
- How the quality assessment and assurance is applied
- Conclusions

Reflecting reality in an LCA database

Massive sustainability database, MSDB, by GreenDelta:

A fresh start for an LCA and sustainability database

- Idea: make a comprehensive, responsive database for LCA and sustainability assessment that reflects reality
- 500,000 datasets, worldwide coverage
- Developed by GreenDelta since 2022
- Sources: wide variety, non-LCA sources, some important specific databases included
- First public review: process to start at openLCA conference, April 2026

Massive sustainability database, MSDB, by GreenDelta:

A fresh start for an LCA and sustainability database

- Idea: make a comprehensive, responsive database for LCA and sustainability assessment that reflects reality

- This is challenging as LCA results cannot be empirically validated

Massive sustainability database, MSDB, by GreenDelta:

A fresh start for an LCA and sustainability database

- Idea: make a comprehensive, responsive database for LCA and sustainability assessment that reflects reality
- Modeling of transportation with parameterised payloads, and empirically validated transportation models (HBEFA, for road transport)
- Modeling of infrastructure with parameterised infrastructure lifetime, capacity, and utilisation
- Use of various datasources, also non-traditional datasources, and apply quality assurance to identify datasets that fit (!)

Transportation modelling in the MSDB

Road transport, use of the full HBEFA model

- Transport effort depends on payload
- Processes formula: effort = offset + payload*factor
- Payload is known at the process using the transportation

  → we need (unfortunately) two processes,
  one that is the offset transportation process,
  one for the factorised transportation effort

- Also cold start processes for short distances

Infrastructure modelling in the MSDB

Machinery and buildings

- Machinery is always included, with parameters:
  - Lifetime
  - Capacity
  - Utilisation

Parameters have default values (process-specific), these can be changed.

Reality in LCA databases:

Was this so important in recent years?

→ **If we want to make LCA datasets and an LCA database that reflects reality*,**

- **we need a different approach for creating it,**
- **and also a quality assurance for the datasets that is somewhat different from quality assurance done today for LCA datasets.**

* In line with LCA methodology, so no storage, linear model

Existing data quality proposals: difficult to apply in dataset creation

- E.g. „precision“, „sample size“ : difficult / impossible to calculate without access to raw data
- „correlation“, „representativeness“: what is the reference?

Table 2. Data quality indicators of various data sources in this study.

	Indicators	Coverage	Scoring	Approach
Ecoinvent	- Reliability - Completeness - Temporal correlation - Geographical correlation - Further technological correlation	Flow level	1 (best)-5 (worst)	Pedigree matrix (e.g. [1;2;2;4;3]) [24,25]
Sphera	- Uncertainty/Precision - Completeness - Time-related representativeness - Geographical representativeness - Technical representativeness - Consistency	Process level	Five levels between “very good” and “very poor”	Expert judgments, ILCD and PEF guidelines [29]
ICE	- Method compatibility - Assurance - Temporal correlation - Geographical compatibility - Transparency - Sample size	Process level	1 (worst)-5 (best)	Average data quality for each category of material [15]
EC3	- Manufacturer specific - Product specific - Facility specific - Batch specific	Process level	True/False	Q-metadata [30]

Gulizar Balcioglu, Amy M. Fitzgerald, Ffion A.M. Rodes, Stephen R. Allen,
Data quality and uncertainty assessment of life cycle inventory data for composites,
Composites Part B: Engineering,
Volume 292,
2025,
112021,
ISSN 1359-8368,
https://doi.org/10.1016/j.compositesb.2024.112021.

The quality assessment in the MSDB

Data quality assessment in the MSDB

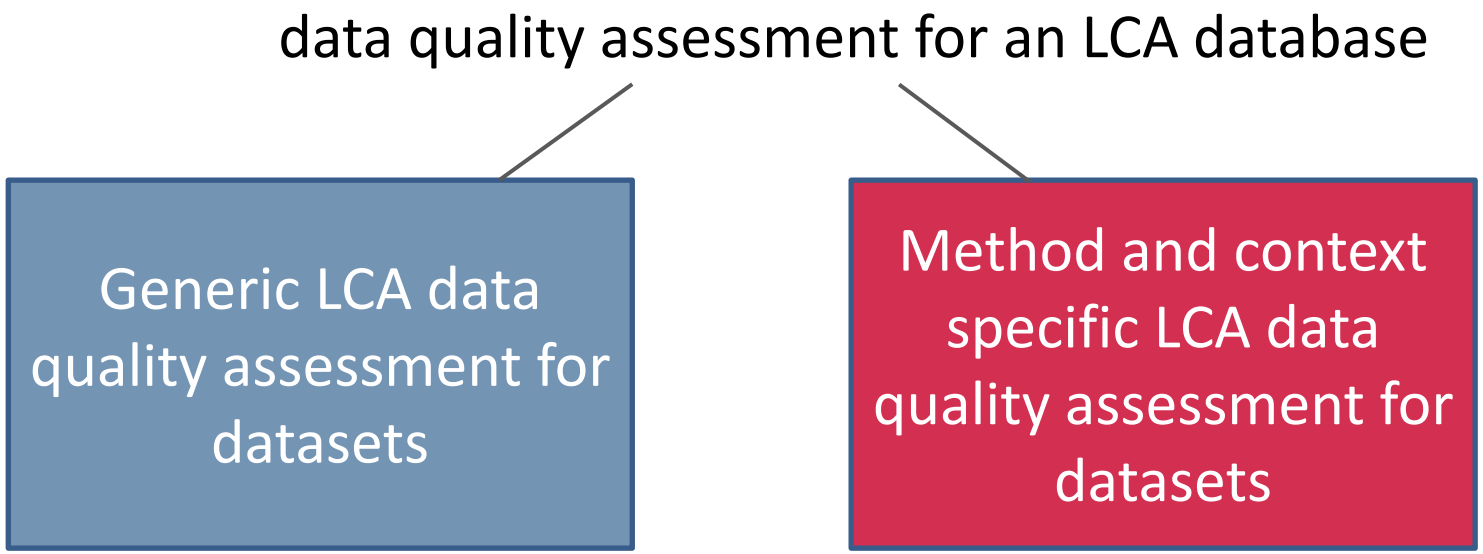
- How can the quality of a dataset be assessed, in a generic database?

  (or in other words, what is a good LCA dataset?)

Data quality assessment in the MSDB

- What is a good LCA dataset?
- How can the quality of a dataset be assessed, in a generic database?

- **Idea: there are two different sets of quality assessments, one independent from specific methodology „flavours“, one reflecting specific context and methodology**



a) Generic LCA data quality assessment

Criteria:

- **Mass & energy balance**
- **Reliability of the source**
- **Process structure**
- **Further logical and natural science rules**

a) Generic LCA data quality assessment

- **Mass & energy balance**

- input is ideally identical to output mass, energy
- Delta mass
  = abs(sum(input_mass) - sum(output_mass)) / average(sum(input_mass) - sum(output_mass))

$$M_{\text{in}} = \sum_{i=1}^{n} m_i^{\text{in}}, \qquad M_{\text{out}} = \sum_{i=1}^{n} m_i^{\text{out}}$$

$$\overline{\Delta M} = \frac{1}{N} \sum_{k=1}^{N} \left(M_{\text{in}}^{(k)} - M_{\text{out}}^{(k)} \right).$$

$$\Delta m = \frac{\left| M_{\text{in}} - M_{\text{out}} \right|}{\overline{\Delta M}}$$

delta mass	score
< 10%	1
< 30%	2
< 50%	3
< 75%	4
< 100%	5
>= 100%	6

a) Generic LCA data quality assessment

- **Reliability of the source**

  - Empirical measurement, confirmed, with source accessible
  - Specific model, with source accessible
  - Publication, with source accessible
  - AI, with source accessible
  - AI
  - Unknown

relaibility of the source	score
Empirical measurement, confirmed, with source accessible	1
Specific model, with source accessible	2
Publication, with source accessible	3
AI, with source accessible	4
AI	5
Unknown	6

a) Generic LCA data quality assessment

- **Process structure**

  - A process has a quantitative reference (it must have a product)
  - The flows in a dataset follow a certain structure
    - Depending on the type of process, the process archetype
    - 20 process archetypes defined, with structures of expected inputs and outputs

6.1. → Verbrennung – Datensatz .....
6.2. → Energiewandlung sonstig – Datensatz .....
6.3. → Stoffwandlung industriell – Datensatz .....
6.4. → Gewinnung Energieträger .....
6.5. → Gewinnung Stoffe Landwirtschaft – Datensatz ...
6.6. → Tierzucht Landwirtschaft – Datensatz .....
6.7. → Gewinnung Stoffe sonstig – Datensatz .....
6.8. → Transport – Strom – Datensatz .....
6.9. → Transport – Pipelines – Datensatz .....
6.10. → Gütertransport .....
6.11. → Personentransport .....
6.12. → Entsorgung .....
6.13. → Mischprozess

Archetype Transformation of materials, industrial – dataset

Definition:

The process transforms one or several input materials into one or several output materials, possibly with occurrence of waste and emissions. The transformation typically requires energy (heat and electricity).

Examples: fused-salt electrolysis, almunium; production of diesel fuel in a refinery.

Inputs:

- starting materials
- probably electricity
- maybe heat
- maybe building (infrastructure)
- maybe flue gas cleaning (infrastructure)
- probably machinery (infrastructure)

Outputs:

- material product
- maybe coproduct
- maybe flue gases
- maybe waste
- maybe waste heat

Process parameters:

- maybe life time of building
- maybe life time flue gas cleaning

Relations within the process

- life time of building parameter depends on building
- life time of flue gas cleaning parameter depends on flue gas cleaning

Other comments:

a) Generic LCA data quality assessment

- **Process structure**

  - Process has a quantitative reference (it must have a product)
  - Flow structure follows archetype structure
    - If we want to model a process of a certain type, we know and can expect a certain structure
    - 20 process archetypes defined, with structures of expected inputs and outputs

structural integrity	score
full fit	1
fit but with few exceptions	2
fit but with some major deviations	3
not fit	4

a) Generic LCA data quality assessment

- **To be met by every dataset, independent from methodology**

- **Further natural science and logic rules**
  - **C in fuel, incinerated-> CO_2 in emission**
  - **Further stochiometric rules**
  - **... (many)**

- **If these are violated, revise the dataset**

b) Method and context specific LCA data quality assessment for datasets

- **Much more straightforward:**
  Meet requirements concerning time, geography, product and process, methodology

- Time difference
- Geography difference
- Technological difference in product and in process
- Methodological compliance

How the quality assessment is applied

The quality assessment is applied:

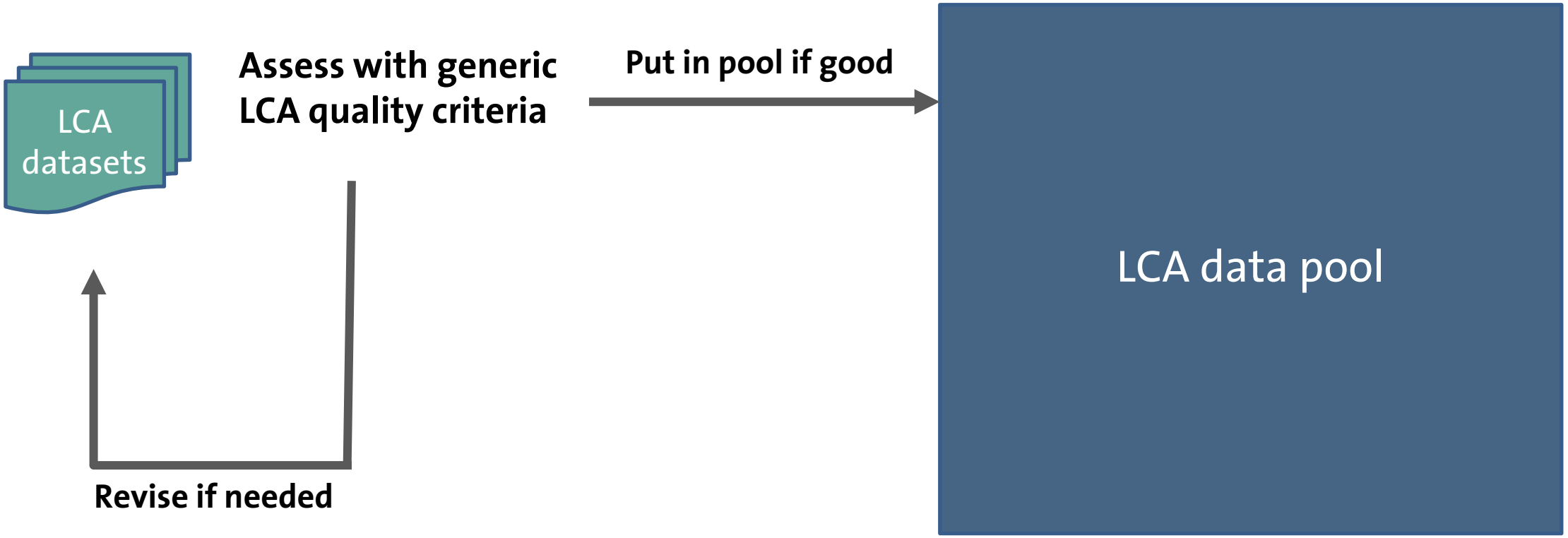
a) Generic LCA quality:

- **As documentation**
- **For the assessment of raw datasets, possibly revision**
- **Find best datasets, when connecting**

b) Specific LCA quality:

- **Meet requirements concerning time, geography, product and process, methodology, when connecting**

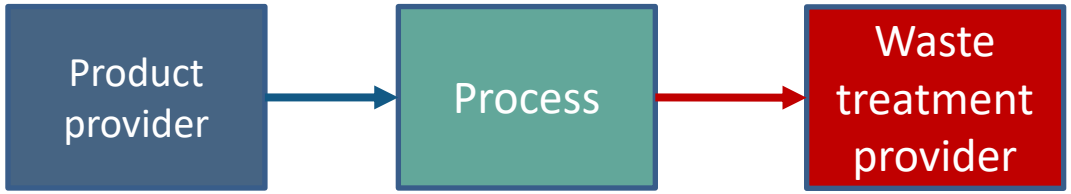
Quality assessment is applied for minimum quality of database datasets



Quality assessment is applied for creating life cycle models, for finding providers

Provider

- = “the next” process that is delivering a product / receiving a waste

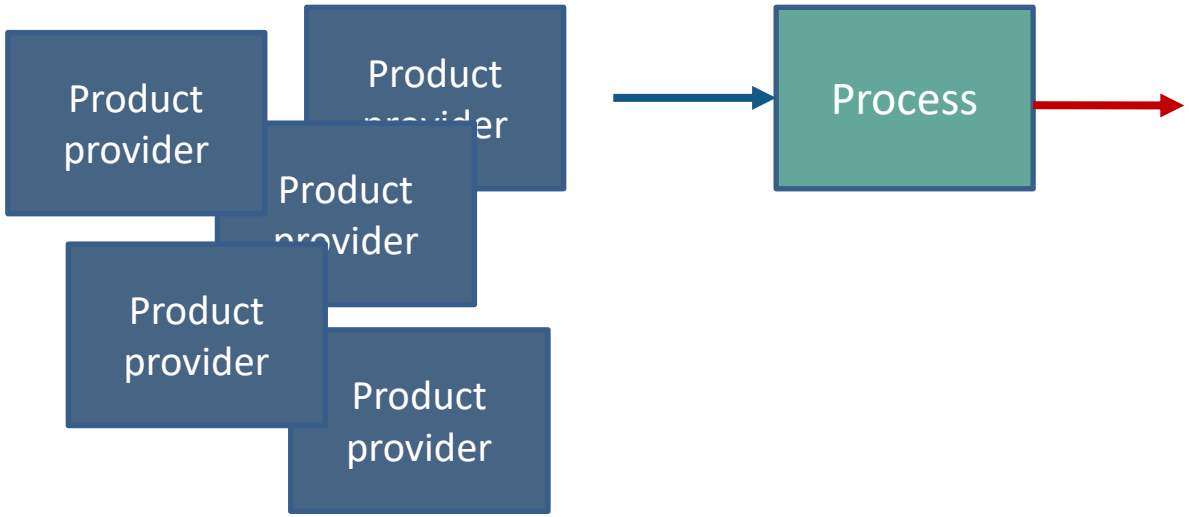


Quality assessment is applied for creating life cycle models, for finding providers

There are ideally several providers

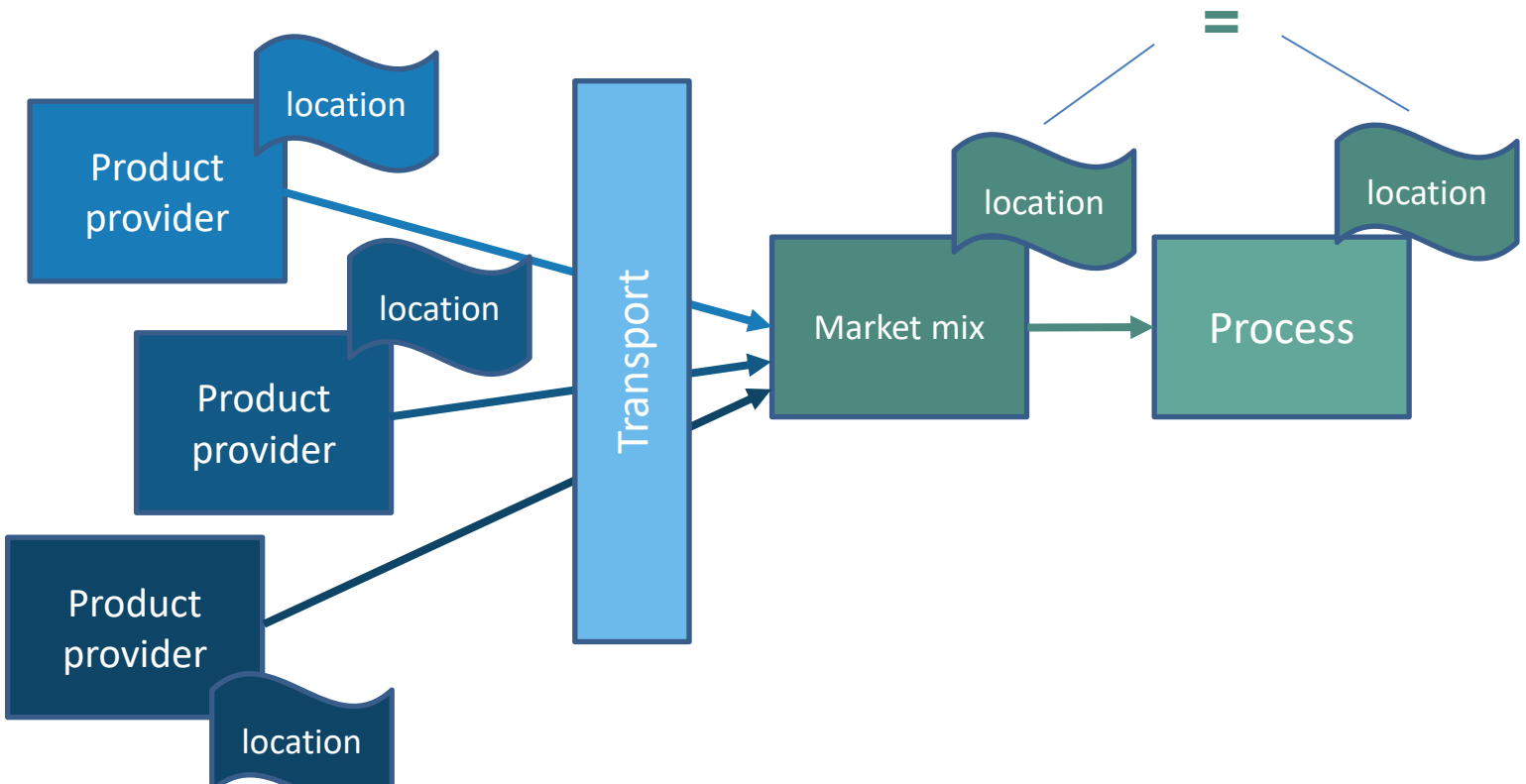
The provider is selected that has the best quality

Generic LCA data quality + method and context specific quality



Quality assessment is applied for creating life cycle models, for finding providers

Transport and market mixes are also considered, for complete life cycle models



GreenDelta

The provider is selected that has the best quality

Generic LCA data quality + method and context specific quality

-> the data pool can contain several datasets for the same purpose, from different data sources

-> the life cycle model can specifically reflect needs of goal and scope -> more specific and better fitting and thus higher quality life cycle models (and results!)

Conclusions and discussion

Conclusions and discussion

1. it is useful to define a data quality that is independent from method selection and goal and scope specific settings of time and geography

2. splitting data quality into generic data quality on the one side and context and method specific data quality on the other side is well suited for a generic database

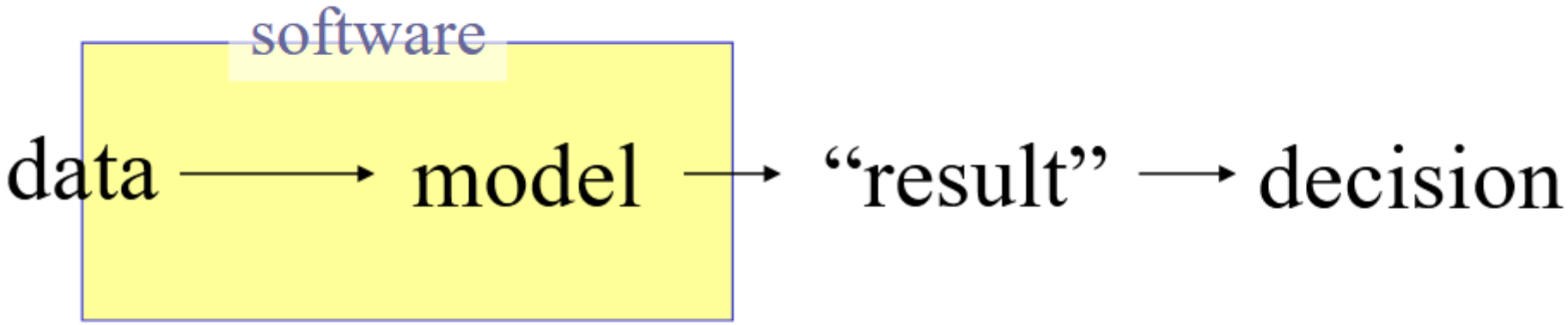
Conclusions and discussion/2

3. In the MSDB, this split data quality is used for an assessment of database raw datasets, and for building life cycle models in the database

4. This flexible creation of life cycle models via providers requires an LCA software that understands providers, and handles life cycle models (such as openLCA)

Conclusions and discussion/3

The role of software in sustainability assessment



Ciroth, A.: A new open source, LCA software , presentation, 7th Ecobalance conference, Tsukuba, presentation and conference transcript, p. 427 ff., November 14th – 16th, 2006.

Finally..

We are starting a public review of the database, with a select group of interested experts.

Please contact us at
gd@greendelta.com,
topic: “review of the MSDB database”

GreenDelta
sustainability consulting + software

Thank you very much!

Dr. Andreas Ciroth, ciroth@greendelta.com

GreenDelta GmbH, Alt-Moabit 130/131, 10557 Berlin, Germany