Freiburg, Germany, March 2019

A General Data Model for Socioeconomic Metabolism and its Implementation in an Industrial Ecology Data Commons Prototype

Stefan PAULIUK,[1,*] *Niko HEEREN,*[2] *Mohammad Mahadi HASAN,*[1] *and Daniel B. MÜLLER*[3]

Journal of Industrial Ecology (2019), DOI: 10.1111/jiec.12890

This is the accepted version of the manuscript, which was later published.

Its content is identical to the content of the journal publication

[1)] Industrial Ecology Freiburg, Faculty of Environment and Natural Resources, University of Freiburg, Germany

[2)] Centre for Industrial Ecology, Yale University, New Haven, Connecticut, USA

[3)] Industrial Ecology Programme and Department for Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[*)] Address correspondence to

Stefan Pauliuk
Industrial Ecology Freiburg
Faculty of Environment and Natural Resources
University of Freiburg, Germany
stefan.pauliuk@indecol.uni-freiburg.de
Tel: +49-761-203-98726

Abstract:

Till this day, data in industrial ecology are commonly seen as existing within the domain of particular methods or models, such as input-output, life cycle assessment, urban metabolism, or material flow analysis data. This artificial division of data into methods contradicts the common phenomena described by those data: the objects and processes in the industrial system, or socioeconomic metabolism. A consequence of this scattered organization of related data across methods is that IE researchers and consultants spend too much time searching for and reformatting data from diverse and incoherent sources, time that could be invested into quality control and analysis of model results instead.

This article outlines a solution to two major barriers to data exchange within industrial ecology: i) the lack of a generic structure for industrial ecology data and ii) the lack of a bespoke platform to exchange industrial ecology datasets.

We present a general data model for socioeconomic metabolism that can be used to structure all data that can be located in the industrial system, including process descriptions, product descriptions, stocks, flows, and coefficients of all kind. We describe a relational database built on the general data model and a user interface to it, both of which are open source and can be implemented by individual researchers, groups, institutions, or the entire community. In the latter case, one could speak of an industrial ecology data commons (IEDC), and we unveil an IEDC prototype containing a diverse set of datasets from the literature.

Keywords: database; socioeconomic metabolism; industrial ecology; open science; OLAP cube;

<heading level 1> **Data integration for industrial ecology research: State and challenges**

Access to data is essential to industrial ecology research. Till this day, data in our field are commonly seen as existing under the domain of particular methods or models, such as input-output, life cycle assessment, urban metabolism, or material flow analysis. Data gathering, labelling, structuring, and organization is method- and model-specific. As a consequence of the scattered organization of related data, Industrial ecology (IE) researchers and consultants spend significant time searching for and reformatting data from diverse and incoherent sources, often with the knowledge that they are not the first ones doing that work for a particular dataset. Time is wasted that could be better used for correct application of methods, quality control, and sensitivity and uncertainty analysis of model results. Data integration and access to data has been a longstanding topic in the IE literature, both in case studies and in theory and database development (Myers et al. 2018; Pauliuk et al. 2016; Lupton and Allwood 2017; Murakami et al. 2010; Frischknecht 2006; Merciai and Schmidt 2018; Weidema 2009; Lenzen et al. 2014; Hertwich et al. 2018; Davis et al. 2015, 2010). For example, the integration of material flow data with input-output (IO) models has a strong tradition (Nakamura and Kondo 2002; Nakamura et al. 2007, 2011; Ohno et al. 2016; Hoekstra and van den Bergh 2006; Merciai and Schmidt 2018). The same holds for data integration into so-called hybrid models that integrate monetary and physical process and supply chain information (Hawkins et al. 2007; Crawford et al. 2018). Still, there is currently no comprehensive inventory of data describing socioeconomic metabolism at different regional, temporal, process, material, and commodity scales and layers such as mass, energy, monetary, or social indicators. Barriers to data integration are both, social and technical (Hertwich et al. 2018):

- Lack of incentives, in particular, funding and data management schemes: When research time is scarce and data management competence is low, inappropriate data archiving and transfer of data into other projects may be the consequence. Research funders are aware of this problem and have started requiring data management plans and open access to data. Scholars from the community have issued a call for greater data transparency in IE research (Hertwich et al. 2018). Still, it is too easy to conduct and publish research that contributes to the knowledge base but not to the cumulative database of the field.

- Lack of cross-method data formats and platforms for exchanging IE data: A general data format for the industrial system does not exist. While generic data sharing platforms such as Figshare (https://figshare.com/) or Zenodo (https://zenodo.org/) have been available for several years and are open and easy to use, they do not prescribe any particular data model or format and have therefore not noticeably alleviated the data integration and exchange problem. (Here, we think of a data model as a scheme that organizes elements of data and specifies how they relate to one another and to real world entities, cf. ecospoldv2 for an example from our community (Meinshausen et al. 2016).)

- A perceived lack of appreciation for sharing data, and the fear of losing competitiveness: Restrictive data sharing policies turn scientific data, a common good, into a local asset. They create local advantages for research groups in a competitive environment. The problem of scientific data turned into local assets is exacerbated by prevalence of consulting-style research in our field. The longer-term benefits of data exchange between individual researchers and the research community may not be appreciated enough.

The different IE research branches have found ways to deal with these barriers, and a number of data integration schemes is available.

<heading level 2> **Currently available schemes and services for data integration**

Highly integrated MRIO (Tukker and Dietzenbacher 2013) and life cycle databases (Wernet et al. 2016) form the informational backbone for entire scientific communities. These databases, however, include only a fraction of the available data on the industrial system needed for the different branches of IE research, and the industrial ecology community as a whole is not backed up by a comprehensive database. In material and energy flow analysis (MEFA) in particular, no common data format and only first attempts towards database development exist, focusing on particular software (www.stan2web.net/), datatypes such as product lifetimes (Murakami et al. 2010), or system scopes such as cities or national economies (Ravalde and Keirstead 2015; Schandl et al. 2017) (https://metabolismofcities.org/). That infrastructure gap leads to inefficiencies across the entire field because the descriptive MEFA forms the basis of process inventorying and other data that feed into both process databases and physical IO tables.

Data pooling services such as Figshare or Zenodo are available for publishing and sharing datasets with low thresholds and thus represent the first and most basic step any researcher should take toward sharing their data with minimal effort. But they haven't alleviated the underlying problem, as these platforms recommend but do not require any structuring of the data, which are often hidden in pdfs or images instead. Nor do these services apply an IE-specific data model needed to provide sufficient metadata and systems context to facilitate transfer of data into other projects.

While the barriers listed above hinder the free exchange of data, there are many datasets available in the literature as well as in different institutional reports. Many government funded IE-related research projects include some type of data-set inventorying as an objective or initial work-package (e.g., http://www.prosumproject.eu/objectives, http://www.mica-project.eu/, or http://www.minea-network.eu/). Still, many published datasets come in custom formatting and often in form of pdfs, figures, or other documents that do not contain structured data. They are not inventoried, nor are they easily accessible by standard modelling tools.

<heading level 2> **Goal and scope of this work**

The goal of this article is to outline a technical solution to two major barriers to data exchange listed above: the lack of a generic data model for IE data and the lack of a bespoke platform to exchange IE datasets. The general data model for socioeconomic metabolism presented below can be used to structure a wide spectrum of data types describing objects (substances, materials, goods, products, or commodities) and processes (industrial transformation, storage, distribution, or consumption) in the industrial system, including process descriptions, product descriptions, stocks, flows, and coefficients of all kind. We describe a relational database built on the general data model and a user interface to it, both of which can be implemented by individual researchers, groups, institutions, or the entire community. In the latter case, one could speak of an industrial ecology data commons (IEDC), and we unveil a prototype for the IEDC that contains a diverse set of published datasets from the literature. The goal of the IEDC is not to encapsulate all IE data and thus replace existing databases, but to offer a data model and a platform for exchanging the many datasets of various types that are scattered across the literature. Future applications in the community and beyond are discussed at the end.

<heading level 1> **A general data model for socioeconomic metabolism**

A candidate for a 'general data model' for socioeconomic metabolism (SEM) must be able to represent the wide spectrum of data describing SEM and its links to the natural environment and to human agents. This spectrum includes: material and product stocks, energy and commodity flows, process yield factors and environmental extensions (emissions and resource use of processes), product lifetimes and material composition, and a large spectrum of performance and impact indicators.

<heading level 2> **Data in a systems context**

Quantitative information on processes, stocks, flows, etc., in socioeconomic metabolism has three components:

1) Value. The actual numerical information, including unit and (optional) uncertainty
2) System location. The information needed to locate the data in the systems context, i.e., the link between data and the system dimensions (process, time, region, material, …)
3) Metadata. Information like provenance, source document, author and version information, and license.

Value and metadata are universal concepts that can be structured using custom or generic methods such as the stats_array system to describe uncertainty information (https://stats-arrays.readthedocs.io/en/latest/) and the Dublin Core Initiative to describe metadata (http://dublincore.org/). The second component, the location of data in the system, is specific to systems science, and needs to be specified in greater detail.

<heading level 2> **System dimensions and data aspects**

Each system definition of socioeconomic metabolism or a subsystem thereof prescribes a number of *dimensions* along which the system content is described: the time dimension is used to order events by the time of their occurrence, the location dimension is used to order objects by their location, the process dimension is used to identify balance volumes to balance or group events, the object (substances, materials, goods, products, or commodities) dimension is used to identify different goods or substances, and the layer dimension is used to indicate the unit in which the data are measured (Table 1).

To locate data in a system definition, one has to specify the *aspects* of the data that describe how the numerical values relate to the time, region, process, etc. dimensions of the system. The data aspects describe how the data relate to the system dimensions. For example, a flow has a starting node and a terminating node. Here, the system dimension 'process' is used to describe two aspects of a flow, namely the starting and terminating process node. A stock is always associated with a node where it is located, and therefore, 'residence process' is an aspect of the 'process' dimension needed to locate a stock. Table 1 lists system dimensions and related aspects.

Table 1. Most common system dimensions for socioeconomic metabolism and the related data aspects. The list of dimensions and aspects is not exhaustive.

System dimension	Description	Related data aspects (example)
Layer	Unit of measurement	Mass, volume, economic value
Process	Transformation, distribution, storage processes	Process of residence (stock), process of origin (flow), process of destination (flow)
Location	Location in space	Region of residence (stock), region of origin (flow), region of destination (flow)
Object (materials and commodities)	Objects of interest (goods, substances, commodities, materials, products, …)	Commodity, good, product group, product type (sub-product), substance, chemical element, waste type, environmental extension
Time	Location in time	Historic time, future (model) time, age-cohort (vintage), time point (stock), time interval (flow)

Scenario	Describing different "realities" or manifestations	Scenario for model drivers, scenario for process parameters

The data model stipulates that each data type (stock, flow, transfer coefficient, lifetime, price, etc.) has a set of core aspects that must be specified to locate the data type in the system and optional aspects that add further specification. For example, to describe the lifetime of a product at least the product and the residence process aspects must be given, and optional aspects include the age-cohort of the product and the region where it is used. Examples for data types and their core and optional aspects are shown in Fig. 1.

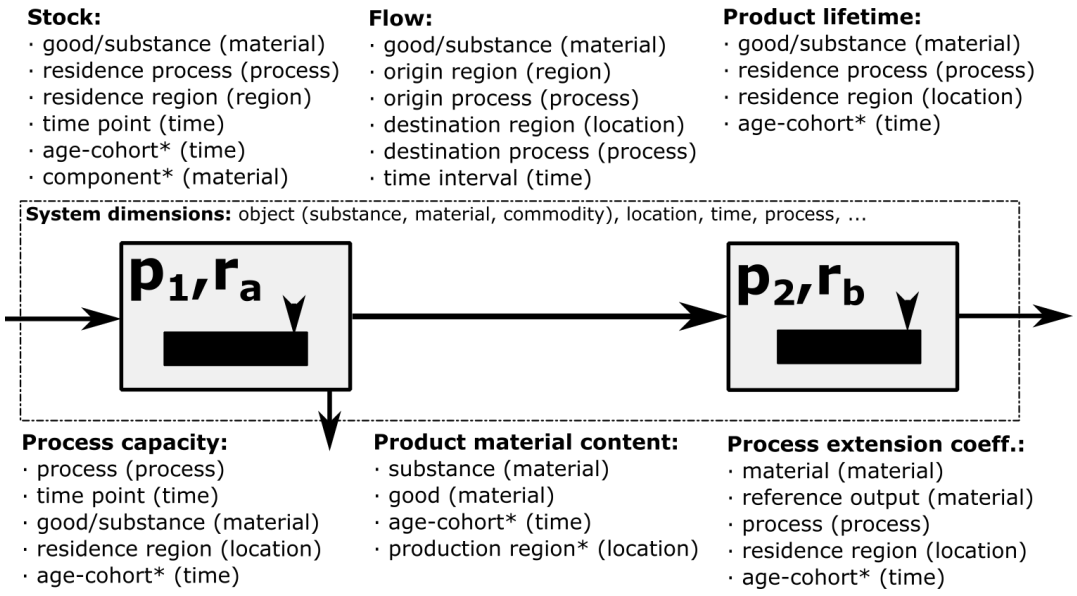

Figure 1. Selected major data types of socioeconomic metabolism, their aspects and respective system dimensions (in brackets). One system dimension, e.g., time, can link to different aspects, e.g., historic time, scenario time, or age-cohort, in the description of the different data types. The system definition shows two processes, p_1 and p_2, in two different regions r_a and r_b, each containing a stock. The asterisk (*) indicates the optional aspects of the different data types. A list of all identified aspects is supplied in Figure S1.

<heading level 2> **Formal definition of the data model**

The location of each data point (number quantifying a fact in a system) in the system can be written as tuple:

$$\begin{aligned} data_item(aspect\ 1,\ aspect\ 2,\ aspect\ 3,\ \ldots) &= value \\ trade_flow(cars, Japan, USA, 2016, number\ of\ units) &= 2540000\ units \end{aligned} \quad (1)$$

Here, a tuple refers to a finite ordered sequence of elements (here: aspects), commonly noted in parentheses (…). The tuple notation of system variables and parameters is commonly applied in IE and other branches of systems analysis, and, notably, for the description of IO contingency tables (Geschke et al. 2011). For a tuple-based data model, each data point is represented as a value assigned to a tuple of data aspects in a discrete multidimensional space. The dimensions of that space represent the different system dimensions (time, region, product, ...), which again are connected to the aspects of the data via the aspect-dimension link (Table 1). The following propositions form the basis of the data model:

(P1) Each data point (numerical value quantifying a fact in a system) requires a certain number of aspects that locate it in the system dimensions.

Only when located in the system's dimensions, the data point has a clear meaning. The next step is to classify data points into types:

(P2) Each data type (family of data describing a certain phenomenon in the system, like stock, flow, material content, product lifetime, …) has a specific data model that i) prescribes which aspects are required and which aspects are optional for the meaningful location of this data type in the system definition, and ii) defines the meaning of each aspect.

We then define the dataset, which comprises a set of data points of a specific data type that are organised together in a certain context, e.g., a figure, table, or model. Examples of datasets are the greenhouse gas emissions of different countries during a given time period (data type:

flow) or the quantities of different materials contained in a certain passenger vehicle type (date type: material composition).

(P3) A socio-metabolic dataset D is a function (or mapping) from a set of aspect domains A_1, A_2, …, A_n into the real numbers, augmented by null for 'no data'. The data type of D prescribes which aspect domains are present, and the elements in the different aspect domains are the values that the different data aspects can take.

$$D : A_1 \times A_2 \times A_3 \times ... \times A_n \rightarrow \mathbb{R} \cup \{\text{null}\} \qquad (2)$$

Each tuple of aspect values locates a data point in the system definition. To describe data uncertainty the individual tuples can be annotated by supplying parameters of probability distributions (for aleatoric uncertainty) and min/max extremes or high-medium-low alternatives to capture epistemic uncertainty. The list of domains A_1, A_2, …, A_n and their meaning is defined by the data type of D (cf. Fig. 1). An aspect domain set contains all possible values for the system dimension describing this aspect, and these values are often referred to as *classifications* (of alloys, products, or industrial sectors). We generalize this term:

(P4) The aspect domain set used to define a dataset is called a classification of the corresponding dimension of this aspect, and the elements in that set, the aspect values, are the classification items.

When handling datasets, not only their data model but also the classifications used for their aspects must be specified. For example, the stock and flow data types shown in Fig. 1 both have a good/substance index/aspect. Aspect domain sets can use custom classifications (custom material 1, custom material 2, custom material 3, ...) or standardized values (year 2000, year 2001, ... or the general product, industry, and substance classifications).

Classifications refer to a specific system dimension, so that different aspects for that dimension can use the same classification.

Finally, we define the data group, which allows us to bundle datasets, e.g., the different stocks and flows in a material cycle description or the flows, production volumes, etc. that make up a unit process inventory:

(P5) A data group is a collection of different datasets from a common source or research project.

Examples and more explanations are provided in the supplementary material.

<heading level 2> **Taxonomy of data types for describing socioeconomic metabolism**

The basic system structure for socioeconomic metabolism consists of processes with stocks and flows between processes (Pauliuk et al. 2016). The two broad categories of description are *objects* such as goods, products, materials or substances, and their storage, distribution, and transformation in *processes*. The objects are described using *properties*, which can be intrinsic, that is independent of system location, e.g., the chemical formula of carbon dioxide, or extrinsic, that means depending on system location, e.g., the magnitude of a trade flow depends on the countries it links. In order to locate an object within the system context, only extrinsic properties must be specified. Further, properties can be intensive (independent of the amount of objects in the system), and extensive (additive for different objects) (Cohen et al. 2008; Pauliuk et al. 2016). The division of system elements into objects and processes and their extrinsic properties into extensive and intensive leads to four general data categories (Table 2), of which one (extensive object properties) can be divided further into stocks and flows, the two basic appearance modes of objects in a system (Pauliuk et al. 2016).

Table 2. The seven general data categories and common data types identified under the data model for socioeconomic metabolism. The list of data types in the table is not exhaustive. All phenomena or properties described in the system are modelled as one of the defined data types. For example, objects of interest appear as stocks or flows, and their intensive properties are modelled as material composition, specific energy consumption, etc.

<table>
<tr><th></th><th>Objects of interest (materials, goods, products, substances, commodities, waste, …)</th><th>Processes (industries, markets, end-use sectors, use-phase)</th></tr>
<tr><td rowspan="2">Extensive (at scale)</td><td>Flows (1)
Flow
Process inventory
Births/Deaths</td><td rowspan="2">Extensive process properties (5)
Output or production capacity of process</td></tr>
<tr><td>Stocks (2)
Stock
In-use stock
Population</td></tr>
<tr><td>Intensive (per unit)</td><td>Intensive object properties (3)
Product material composition
Product lifetime
Price of products
Specific energy consumption</td><td>Intensive process properties (4)
Process yield factors
Process environmental extensions per output
Process operating costs per output
Unit process inventory</td></tr>
<tr><td></td><td colspan="2">General ratios (6)
Per capita stock
Per capita flow
Material substitution coefficient
Impact indicators</td></tr>
<tr><td></td><td colspan="2">Correspondence (7)
Correspondence tables between classifications</td></tr>
</table>

Data for individual or groups of objects, such as product lifetime or product material composition, describe the intensive object properties. The analogous group for processes contains, for example, the yield ratios of manufacturing processes, the greenhouse gas emissions per unit of output, or operating costs per unit of output. Moreover, data of categories 1-5 can be used to define ratios, like GDP per capita or per capita building stock, which together form a sixth data category ‘general ratios’. Finally, a seventh data category was created to store correspondences between different classifications. While the list of seven

data categories is a core part of the general data model, the definition of specific data types under these categories the result of consensus-building among data providers.

<heading level 2> **Examples of data types and their aspects in the data model.**

To show how the data model applies to different data types, we list a proposal for the required and optional aspects of a number of frequently used data types, and provide a definition for each type, which describes how the numerical information [value/uncertainty] for a given mass, volume, energy, or monetary [layer] relates to the different aspects of the dataset. Aspects with (*) are optional.

- **Flows (category 1):** A flow is an extensive system variable describing the relocation of material (good or substance): [Flow] of [material] from [origin_process] in [origin_region] to [destination_process] in [destination_region] in [time] period is [value/uncertainty] for [layer].

- **Stocks (category 2):** A stock is an extensive system variable describing the location of material (good or substance): [Stock] of [material] of [age-cohort*] in [process] in [region] of location at [time] point is [value/uncertainty] for [layer].

- **Material composition of products (category 3):** The material composition is an intensive object property describing the proportion of a material in a good/substance: [Material content] of [material] in [good/substance] of [age-cohort*] in production [region*] is [value/uncertainty] for [layer]. The layer can be mass but also volume.

- **Product lifetime (category 3):** The lifetime is an intensive object property describing the residence time of a material (good/substance) in a process: [Lifetime] of [material] of [age-cohort*] in [process] in [region] is [value/uncertainty] for [layer].

- **Process yield factors (category 4):** The process yield factor is an intensive process property describing the share of a material in an input good/substance that is transformed or manufactured into an output good/substance: [Yield] of [material] in [input_commodity] into [output_commodity] in [process] of process [technology*] of [age-cohort*] in [region] is [value/uncertainty] for [layer]. The layer can be mass but also volume.

- **Process extension ratios (category 4):** The process extension is an intensive process property describing the amount of a material flowing from or to the environment per output commodity in a certain process: [Extension coefficient] of [material] per [output_commodity] in [process] of process [technology*] of [age-cohort*] in [region] in [time] is [value/uncertainty] for [layer]. Many different layers are possible: mass per mass, mass per value, volume per value, etc.

- **Process capacity (category 5):** The process capacity is an extensive process property describing the maximum rate of output commodity that can be produced in a certain process: [Capacity] of producing [output_commodity] in [process] of [technology*] of [age-cohort*] in [region] in [time] is [value/uncertainty] for [layer]. The layer can be mass, number of items, or monetary value.

- **Per capita stock (category 6):** The per capita stock is a general ratio describing the amount of a commodity per person in a certain process and region: [Per capita stock] of [material] of [age-cohort*] in [process] in [region] at [time] point is [value/uncertainty] for [layer].

A list of all hitherto defined datatypes can be found in the supplementary material and on the IEDC GitHub repository.

<heading level 2> **Structure and resolution of different data types**

The explicit listing of the core and optional aspects for each data type, introduced by the data model presented here, allows scholars from different modeling communities to enter a dialogue and reach consensus about the aspect structure and the semantics of the different data types that are used across methods. The data model imposes a common structure for the different data types. But it remains flexible regarding system scope and resolution, the choice of which remains research question-driven, project-specific, and at the discretion of the data gatherers. For each dataset the classification items of the different aspects can follow an established global classification, such as ISO regional codes, use a project-wide classification such as the ecoinvent 3.4 activity list (https://www.ecoinvent.org/), or can be given in a custom classification that applies to a given dataset only. Intrinsic information about the different aspect items, i.e., information that is independent of the system context (chemical formula of CO_2, region ISO codes, ...), can be recorded as separate attributes along with the dimensional classification items.

<heading level 1> **Implementation of the general data model in the industrial ecology data commons**

The data model can be implemented in different ways, including spreadsheet-formatted data, relational databases, or array-shaped data in programming environments. Data accessibility and the level of database integration determine the circle of potential users and applications and require proper consideration. Here, we present a proposal for an Industrial Ecology Data Commons (IEDC) based on the general data model.

<heading level 2> **Choice of database integration level.**

A number of design choices needs to be made when developing a database for socioeconomic metabolism; they fall into the following categories:

- Data structure: Extent to which a specific data model is prescribed.
- Data aspect classification: Custom, project-specific, discipline-specific, or universal
- Data interoperability: Level of formatting and machine-readability

Different levels of data integration are possible, and the following three broad cases are typical (cf. also Table 3):

- Low level of integration: Data are inventoried by keywords but come in custom aspect classification and file format. Figshare, Zenodo, or CKAN-based data libraries are examples (https://ckan.org/).
- Intermediate level of integration: Data are cast into data models but not integrated as the different datasets keep their original classifications. Data are easy to access to be used for a wide range of research questions. This is the gap the IEDC tries to fill.

- High level of integration. Data, also of different types, are interlinked via a project-wide classification for the different data aspects, which makes the database internally consistent. MRIO tables, life cycle databases, or UN trade statistics are examples of different degrees of data linking. Large datasets are relatively easy to use but the set of research questions they are applicable to is limited due to given classification and data scope.

Table 3. Properties of different database integration levels. The low, intermediate, and high integration levels were singled out for the purpose of creating this overview, they represent a spectrum of many different nuanced integration levels.

Integration level	**Semantics/ structure**	**Aspect classification**	**Accessibility**	**Data linking**
Low, e.g. Zenodo	No data model used	Custom classifications prevail	All data formats are allowed, including figures and pdfs.	No linkages, possibly shared classifications without harmonized identifiers.
Intermediate, e.g. IEDC (this work)	Common data model used	Mix of general and custom classifications	All datasets are machine-readable in common format	Thematic linkages, some shared classifications with harmonized and unique identifiers.
High, e.g. ecoinvent	Common semantic model across all datasets	Common classification used across all datasets	All datasets are machine-readable in common format	Fully linked datasets with shared classification and unique identifiers

Other integration levels can be thought of. For example low integration with the additional requirement that all datasets are version-managed and machine-readable, such as with the *Dat Project* (https://datproject.org/).

<heading level 2> **Database structure of the industrial ecology data commons**

For the industrial ecology data commons (IEDC) prototype we chose the intermediate level of integration because of the prevailing infrastructure gap. The common data model presented above is used, a mix of general, subject-specific, and custom aspect classifications is allowed for and all datasets are transformed into a common data format. All data can be linked to general classifications, which again can be linked to the Semantic Web. The tuple-based data model presented above is represented as OLAP cube, where OLAP stands for 'online analytical processing'. An OLAP cube is a multi-dimensional array representation of tuple-based datasets, also applicable to IE data (Lupton and Allwood 2017; Gray et al. 1996). In applications and databases, OLAP cubes can be implemented in different ways, e.g., multidimensional arrays, 2D tables with multi-indices, or a list of tuples.

We chose a relational database model for implementing the OLAP cubes as it is well established, easy to set up and therefore suitable for a prototype. A simplified representation of the relational database structure is shown in Figure 2. The database core comprises six tables (*data* for storing data points, *datasets* for storing dataset aspect structure and metadata, the *projects* and *datagroups* tables to group datasets, and the *classification_definitions* and *classification_items* tables to define and describe the dimensional classifications used). The IEDC does not contain one but many different datasets, which means that next to a data table, we need a dataset table describing the meaning and grouping of the different data table entries. While the data points are linked to individual classification items, the dataset table points to one specific classification for each aspect. Thus, classification definitions and classification items must be separate tables. Finally, the aspects link to system dimensions, and the data types link to the data categories 1-7.

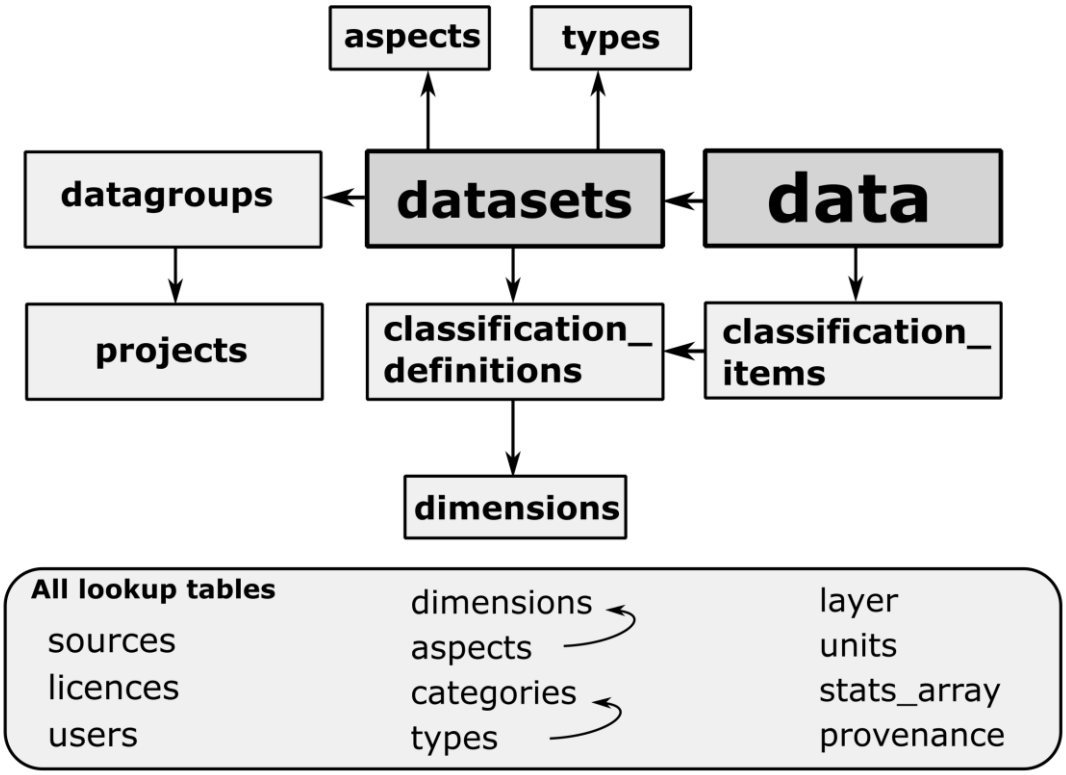

Figure 2. Relational database model for the industrial ecology data commons (IEDC). Each rectangular box represents a table, the arrows the relations (foreign keys) between tables. The relations to most of the 11 lookup tables are not shown to focus on the links between core tables. The complete table description, the list of foreign keys, and the database creation code is available on the project's GitHub repository and in the supplementary material.

<heading level 2> **Implementation of the prototype**

For reasons of convenience, a MySQL relational database server was chosen because it is easy to set up and maintain, also by non-experts, and graphical user interfaces are available from off-the-shelf libraries. An Excel template featuring both table and list-shaped data and a Python import library (https://github.com/IndEcol/IEDC_tools) are used to transfer template data into the database. Data providers need to apply basic formatting to incoming data and make an effort to use already existing classifications to label the different data aspects. The data importer will then automatically link the supplied individual classification items of the different data aspects to the already defined classifications. If a custom classification is used

for a certain aspect this is indicated in the data set and a new classification is then created and used during the upload routine. Details are provided in the supplementary material. For non-template datasets custom scripts can be used as well. Uncertainty information for random and systematic errors can be provided for each data point. There are options for entering the parameters for the commonly used probability distributions as well as low-mean-high alternatives or standard deviations.

The data importer has the following functional specification: (i) parse data template and read all metadata and data and make sure that all are formatted correctly; (ii) check whether all metadata (user, data type, data category) exist in the database so that the new data can be linked to them; (iii) check whether classification items for already registered classifications exist so that the new data can be linked to them; (iv) create custom classifications from classification items gathered from data template so that the new data can be linked to them; (v) link new data to classifications and lookup tables and insert them into database.

For quality control purposes, data are first uploaded to the review database, which has the same structure as the main database. Only once the data provider has confirmed that data were uploaded correctly, the dataset is moved to the main database. Figure 3 shows the data flow for uploading, reviewing, and retrieving data. While the web interface can be used to search for, select, and download individual datasets in different formats, bulk upload and download as well as custom querying of the database are available by means of SQL queries. Built-in or external Object-Relational Mapping packages (ORMs) facilitate direct access from e.g., Python (pymysql) or Matlab.

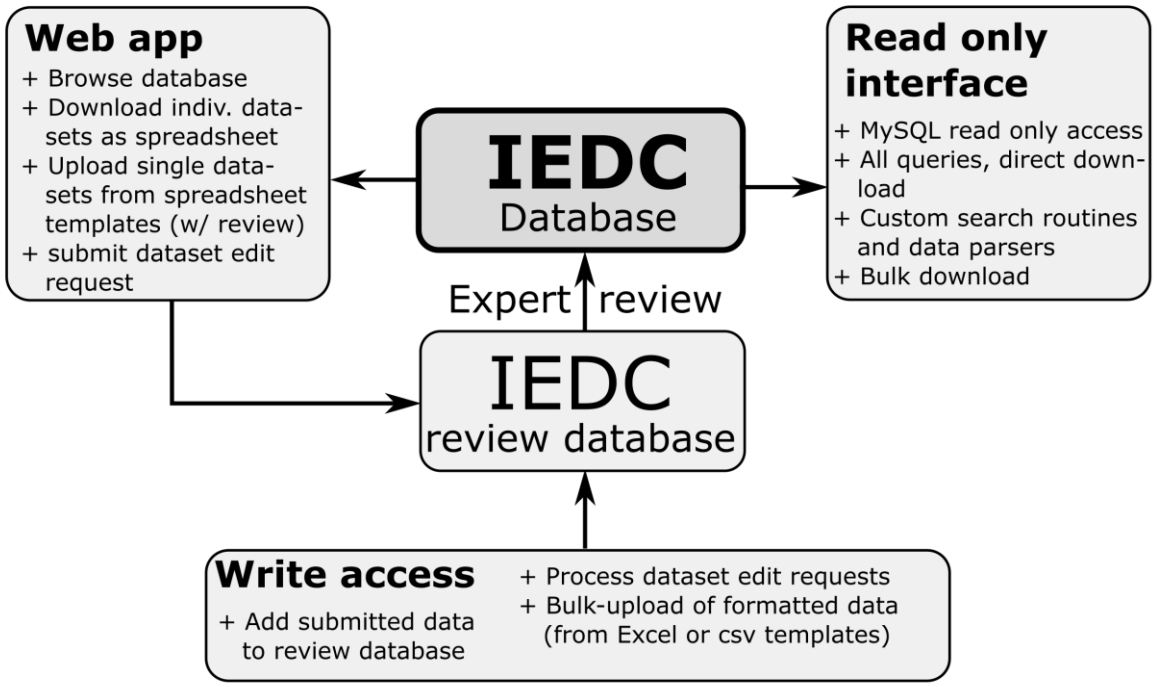

Figure 3. Proposed industrial ecology data commons data flow for uploading, retrieving and downloading data.

The IEDC prototype presented here allows scholars to record the system location and the meaning of a wide array of socio-metabolic data in a systems context. The data model does not record and cannot resolve conflicts between different datasets that result from classification conflicts or unclear or conflicting system boundary settings. Resolving such conflicts requires additional work, and the resulting refined datasets can again be distributed via the IEDC. To prepare the data for future use in other projects, data suppliers should make an effort to document background information, especially where possible data interpretation conflicts are known to arise, e.g., for ore extraction (before or after concentration?) or mass and energy layers (dry or wet mass? Lower or higher heating value?). The prototype of the Industrial Ecology Data commons, containing more than 800 000 data points in 128 datasets on 2019-02-15, is accessible via www.database.industrialecology.uni-freiburg.de.

<heading level 1> **Data integration for industrial ecology research: The way forward**

The data model and data commons prototype presented here offer an immediate benefit to the community by providing a platform for direct download of available datasets and by offering a method for structuring datasets that currently hibernate in pdfs, spreadsheets, and other documents. Below, we present our initial thoughts on the wider implications of the data model.

<heading level 2> **Linking data across industrial ecology and beyond**

Translation schemes to convert established data models to the comprehensive one presented here are already available on the IEDC GitHub repository, including the *ecospold* activity data model (Meinshausen et al. 2016), the general knowledge model for LCA (Kuczenski et al. 2016; Kuczenski 2018), and the tabular representations of MRIO data for which a tuple-based data model already exists (Geschke et al. 2011). Translations from other available formats such as the Sankey diagram (Schmidt 2008) data format (Lupton and Allwood 2017), the unified materials information system (Myers et al. 2018), the data format of the STAN software for MFA (Brunner and Rechberger 2016), a recently published data characterization framework for MFA (Schwab et al. 2016), and the databases for urban metabolism (https://metabolismofcities.org/), (Ravalde and Keirstead 2015) would represent useful future additions.

Seeing the format of established databases such as ecoinvent datasets and MRIO tables as different representations of a common underlying structure helps to understand and compare the different data models used, which is crucial when combining databases in a more automated manner than it is done today. Complex datasets, such as an ecospold unit process

descriptions or MRIO tables, can be broken down in the more basic data types of the IEDC. E.g., an IO table consists of different flow and coefficient tables (inter-industry, industry to environment, industry to final consumers, …), and an LCA unit process inventory (ecospold) consists of normalized flow data, production volumes, and intensive data such as water content and price information. All these individual datasets fit into the general data model for SEM presented here, and their combination into a subject-specific database can be described as a data group.

Once the data have been recast into a general structure, they can be linked to each other to speed up the extraction and analysis of datasets describing entire systems, for example, to quickly find all data related to 'steel production' in 'China' in '2015', no matter which database, method, or subfield they come from. In a first step, datasets can be linked by supplying common keywords. Exact (and machine-operable) links, however, can only be established by using common classifications for the different data aspects across datasets.

To improve the usefulness of the data outside industrial ecology and link them to other information, explicit links to other data realms need to be created. This can be done by linking the common classifications used in the IEDC to the Semantic Web by entering the Uniform Resource Identifiers (URI) of generally used classification items such as chemical elements, countries, or material groups into the *classification_items* table, thus creating a pool of linked open data (LOD) (Ghali and Frayret 2018; Davis et al. 2010). If data providers use URIs as well, automatic data updates could become a standard procedure in the future. To further facilitate machine-processing of data, the linked IEDC data could be broken down into *subject-predicate-object* triples, commonly known as Resource Description Framework (RDF) data (Schreiber and Raimond 2014). A pilot project on using RDF for storing IE data conducted by the main author found that the toolchain for storing and querying triple

databases was much harder to implement and operate than the off-the-shelf relational database servers, so that the latter remain the more practical option given the presently available resources.

<heading level 2> **Expansion and limits of the data model**

The tuple notation of data is very general and flexible due to the arbitrary number of dimensions, aspects, and classifications allowed. Footprints and criticality metrics, for examples, are systems context indicators that cannot be measured directly and that are qualitatively different from process indicators and intensive object properties like material content. Nevertheless, they can be recorded under the data model, simply by defining a 'footprint' or 'criticality' aspect and recording the corresponding data for different regional, temporal, and system scopes.

MEFA is often used to describe processes in the environment (e.g., nutrient balances in the soil, in lakes, or forests) and the data model presented here can be applied to non-industrial systems as well, as long as suitable system dimensions and data aspects are chosen.

The data model and the IEDC were created from the necessity to describe the system in discrete regions, processes, and commodity groups. For geographic data, such as those store in shapefiles and other data that describe a continuum, such as high resolution time series or satellite images, the data model still applies as long as these data can be placed in a system definition of the type described by Pauliuk et al. (2016). The database structure presented here cannot efficiently accommodate very large datasets, such as high-resolution data, however.

The data model allows for recording of basic metadata at the data item and dataset level. Metadata specification varies substantially across fields, (e.g., compare available ecospold

metadata with available MRIO metadata) and future work needs to identify how these rich but differently formatted metadata can be brought into a general structure.

<heading level 2> **Building the foundation of community data infrastructure**

The IEDC prototype is a demonstrator for how data infrastructure in the industrial ecology community could function at intermediate levels of data integration and accessibility. It fills the gap between data pools such as Figshare and Zenodo and the highly integrated life cycle and MRIO databases. It offers new functionalities to IE researchers and can inform the debate on what type of data infrastructure the IE community should invest in.

Next to a common underlying data structure, the main barriers to higher levels of data integration are that universal classifications for processes, products, materials, etc. are not commonly used, and existing classifications are not linked to the Semantic Web by entering the Uniform Resource Identifiers (URI). While undesirable from a data exchange and linking perspective, the flexible use of classification facilitates research in case studies as it eliminates the overhead of having to conform to a predefined classification. The customization of both system definition and classifications used is a basic principle of MEFA research, in particular. Because of their flexibility, the work with custom classifications will persist in the future and the community data infrastructure needs to accommodate this need: Use general classifications and linked data wherever possible, allow for custom classifications where necessary. Precise recording and description of custom classifications (e.g., points of measurement, composition of product and material categories, time frames of measurement) will help data users to interpret the data and aggregate, disaggregate, or use them as proxies in subsequent work. Such frameworks for data collection are being developed for MEFA

(https://minfuture.eu/results), (Fischer-Kowalski et al. 2011), LCA (EU JRC 2010), and for IO, the System of National Accounts already exists (UN 2008).

<heading level 2> **Review, version control and traceability of data**

The IEDC prototype stores data that are published elsewhere, so that these data have already passed quality control like peer review, and the main focus of the review process when uploading to IEDC lies on the correct application of the data model chosen and the correct transfer of numbers. This formatting process and the subsequent use of data often leads to insights that warrant the publishing of a revised dataset version, and the IEDC data structure facilitates the recording of different versions of a dataset due to the UNIQUE constraint on the (dataset_name; dataset_version) tuple in the dataset table. This setup does indeed not allow for the versioning of individual data items within a dataset. Here, the link between data version management tools (e.g., via https://datproject.org/) and a static database needs to be explored. The IEDC prototype offers upload and download options from and to Excel templates (containing the dataset description and the data items) and from and to custom formats using scripts using SQL queries. More versatile formats need to be developed, such as JSON, rdf, or csv files, and shared and version-managed as so-called data packages (https://frictionlessdata.io/data-packages). For the time being, the documentation of the process of converting available data from the different sources, accounting routines, or model calculations into the IEDC template is the under the responsibility of the data provider, and no guidelines or standards exist. A convenient way of tracing this process is to use additional sheets in the Excel template. For data that are only available in a non-portable format, such as tables or figure in image or PDF files, we encourage the use of the liberated_data project

(https://github.com/nheeren/liberated_data) to increase reproducibility and transparency in the data extraction process.

<heading level 2> **Licensing and legal issues**

When compiling the data for the prototype we realized that most published datasets don't specify a license for their use. Some appear to 'inherit' the license of the document containing them, such as the supplementary material of journal publication. The lack of licensing transparency as well as the barriers of re-using data due to intended or unintended restrictive licensing clearly are problems that need to be solved. The issue has already been taken up by a recent call for more data transparency in IE (Hertwich et al. 2018) and by the open energy system modelling community, who, in a recent review, recommend regarding the choice of licenses for datasets: "In the case of data and other content, common permissive licences are Creative Commons Attribution licences (CC BY) and common copyleft licences include other Creative Commons licences such as Creative Commons Attribution-ShareAlike (CC BYSA) or the Open Data Commons Open Database License (ODbL)" (Pfenninger et al. 2017). Pfenninger et al. (2017) also state the importance of choosing a license, as data and code that is provided without a license falls under copyright protection by default and cannot legally be used by others.

<heading level 1> **Conclusion**

The data model for socioeconomic metabolism provides a general structure for many common data types, including stocks, flows, concentrations, lifetimes, and process coefficients. It clarifies the relationship between data, the underlying data types and categories, the data aspects, and the link between data aspects and system dimensions. Well-structured, well-described, and accessible data are key to storing data for later use, both within and across research groups. The data model presented here helps structure and consolidate quantitative systems information and thus make data easier to archive and re-use without loss of meaning. Moreover, the possibility of interlinking data allows for new ways of industrial ecology systems analysis.

A data commons is a major building block for open science in industrial ecology. A strong and vibrant research community, fueled by a rich and open database, is attractive to future scholars, visible to other research fields, and has high impact with decision makers and the general public. Its establishment requires commitment by individual researchers to submit data and by the community to build up and maintain the infrastructure. The benefit of such an investment would be immense.

Supplementary material

A summary of the database structure and the hitherto defined data types is available as online supplementary material. The entire project material is available on GitHub under an MIT license via https://github.com/IndEcol/IE_data_commons/. The prototype of the Industrial Ecology Data commons (IEDC), containing more than 800 000 data points in 128 datasets on 2019-02-15 published under different licenses, is accessible via www.database.industrialecology.uni-freiburg.de

Acknowledgement

The authors thank the members of the Data Transparency Task Force of the International Society for Industrial Ecology, in particular, Brandon Kuczenski, Rick Lupton, and Konstantin Stadler, for providing feedback on the data model and for pointing us to additional references. Thomas Millross helped us clarify the presentation of the data model and the database structure. Rio Aryapratama, Moritz Bisch, Thibaud Pereira, and Paula Vollmer extracted and formatted datasets from the literature for filling and testing the prototype.

References

Brunner, P.H. and H. Rechberger. 2016. *Practical Handbook of Material Flow Analysis. 2nd edition*. CRC Press.

Cohen, E.R., T. Cvitas, J.G. Frey, B. Holmström, K. Kuchitsu, R. Marquardt, I. Mills, et al. 2008. *Quantities, Units and Symbols in Physical Chemistry", IUPAC Green Book, 3rd Edition*. Cambridge: IUPAC & RSC Publishing.

Crawford, R.H., P.-A. Bontinck, A. Stephan, T.O. Wiedmann, and M. Yu. 2018. Hybrid life cycle inventory methods - A review. *Journal of Cleaner Production* 172: 1273–1288.

Davis, C., I. Nikolic, and G.P.J. Dijkema. 2010. Industrial Ecology 2.0. *Journal of Industrial Ecology* 14(5): 707–726. http://doi.wiley.com/10.1111/j.1530-9290.2010.00281.x. Accessed October 13, 2014.

Davis, C.B., A. Chmieliauskas, G.P.J. Dijkema, and I. Nikolic. 2015. *Enipedia*. Delft, The Netherlands. http://enipedia.tudelft.nl/.

EU JRC. 2010. *International Reference Life Cycle Data System (ILCD) Handbook - General guide for Life Cycle Assessment*. Luxembourg: European Union.

Fischer-Kowalski, M., F. Krausmann, S. Giljum, S. Lutter, A. Mayer, S. Bringezu, Y. Moriguichi, H. Schütz, H. Schandl, and H. Weisz. 2011. Methodology and Indicators of Economy-wide Material Flow Accounting. *Journal of Industrial Ecology* 15(6): 855–876.

Frischknecht, R. 2006. Notions on the design and use of an ideal regional or global LCA database. *International Journal of Life Cycle Assessment* 11: 40–48.

Geschke, A., M. Lenzen, K. Kanemoto, and D.D. Moran. 2011. AISHA : A tool to construct a series of contingency tables. In *19th IIOA Conference*, 1–61.

Ghali, M.R. and J.-M. Frayret. 2018. Social Semantic Web Framework for Industrial Synergies Initiation. *Journal of Industrial Ecology* 0(0). https://doi.org/10.1111/jiec.12814.

Gray, J., A. Bosworth, A. Lyaman, and H. Pirahesh. 1996. Data cube: a relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS. In *Proceedings of the Twelfth International Conference on Data Engineering*, 152–159.

Hawkins, T.R., C. Hendrickson, C. Higgins, H.S. Matthews, and S. Suh. 2007. A mixed-unit input-output model for environmental life-cycle assessment and material flow analysis. *Environmental Science & Technology* 41(3): 1024–31.

Hertwich, E.G., N. Heeren, B. Kuczenski, G. Majeau-Bettez, R.J. Myers, S. Pauliuk, K. Stadler, and R. Lifset. 2018. Nullius in Verba. Advancing Data Transparency in Industrial Ecology. *Journal of Industrial Ecology* 22(1): 6–17.

Hoekstra, R. and J.C.J.M. van den Bergh. 2006. Constructing physical input-output tables for environmental modeling and accounting: Framework and illustrations. *Ecological Economics* 59(3): 375–393.

Kuczenski, B. 2018. Disclosure of Product System Models in Life Cycle Assessment: Achieving Transparency and Privacy. *Journal of Industrial Ecology* 0(0). https://doi.org/10.1111/jiec.12810.

Kuczenski, B., C.B. Davis, B. Rivela, and K. Janowicz. 2016. Semantic catalogs for life cycle assessment data.

Journal of Cleaner Production 137: 1109–1117. http://dx.doi.org/10.1016/j.jclepro.2016.07.216.
Lenzen, M., A. Geschke, T.O. Wiedmann, J. Lane, N. Anderson, T. Baynes, J. Boland, et al. 2014. Compiling and using input-output frameworks through collaborative virtual laboratories. *The Science of the Total Environment* 485–486C: 241–251.
Lupton, R.C. and J.M. Allwood. 2017. Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. *Resources, Conservation & Recycling* 124(October 2016): 141–151. http://dx.doi.org/10.1016/j.resconrec.2017.05.002.
Meinshausen, I., P. Müller-Beilschmidt, and T. Viere. 2016. The EcoSpold 2 format—why a new format? *The International Journal of Life Cycle Assessment* 21(9): 1231–1235. http://link.springer.com/10.1007/s11367-014-0789-z. Accessed August 22, 2014.
Merciai, S. and J.H. Schmidt. 2018. Methodology for the Construction of Global Multi-Regional Hybrid Supply and Use Tables for the EXIOBASE v3 Database. *Journal of Industrial Ecology* 22(3): 516–531.
Murakami, S., M. Oguchi, T. Tasaki, I. Daigo, and S. Hashimoto. 2010. Lifespan of Commodities , Part I - The Creation of a Database and Its Review. *Journal of Industrial Ecology* 14(4): 598–612.
Myers, R.J., T. Fishman, B.K. Reck, and T.E. Graedel. 2018. Unified Materials Information System (UMIS): An Integrated Material Stocks and Flows Data Structure. *Journal of Industrial Ecology* in press.
Nakamura, S. and Y. Kondo. 2002. Input-Output Analysis of Waste Management. *J. Ind. Ecol.* 6(1): 39–63.
Nakamura, S., Y. Kondo, K. Matsubae, K. Nakajima, and T. Nagasaka. 2011. UPIOM: a new tool of MFA and its application to the flow of iron and steel associated with car production. *Environmental Science & Technology* 45(3): 1114–20.
Nakamura, S., K. Nakajima, Y. Kondo, and T. Nagasaka. 2007. The Waste Input-Output Approach to Materials Flow Analysis Concepts and Application to Base Metals. *Journal of Industrial Ecology* 11(4): 50–63.
Ohno, H., P. Nuss, W.-Q. Chen, and T.E. Graedel. 2016. Deriving the Metal and Alloy Networks of Modern Technology. *Environmental Science and Technology* 50(7): 4082–4090.
Pauliuk, S., G. Majeau-Bettez, E.G. Hertwich, and D.B. Müller. 2016. Toward a Practical Ontology for Socioeconomic Metabolism. *Journal of Industrial Ecology* 20(6): 1260–1272.
Pfenninger, S., L. Hirth, I. Schlecht, E. Schmid, F. Wiese, T. Brown, C. Davis, et al. 2017. Opening the black box of energy modelling : Strategies and lessons learned: 1–9.
Ravalde, T. and J. Keirstead. 2015. A database to facilitate a process-oriented approach to urban metabolism. *Journal of Industrial Ecology* accepted(0): 1–12.
Schandl, H., M. Fischer-Kowalski, J. West, S. Giljum, M. Dittrich, N. Eisenmenger, A. Geschke, et al. 2017. Global Material Flows and Resource Productivity: Forty Years of Evidence. *Journal of Industrial Ecology* 22(4): 827–838. https://doi.org/10.1111/jiec.12626.
Schmidt, M. 2008. The Sankey Diagram in Energy and Material Flow Management. Part I: History. *Journal of Industrial Ecology* 12(1): 82–94. http://doi.wiley.com/10.1111/j.1530-9290.2008.00004.x. Accessed June 1, 2015.
Schreiber, G. and Y. Raimond. 2014. RDF 1.1 Primer. https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/. Accessed January 24, 2019.
Schwab, O., O. Zoboli, and H. Rechberger. 2016. A Data Characterization Framework for Material Flow Analysis. *Journal of Industrial Ecology* 21(1).
Tukker, A. and E. Dietzenbacher. 2013. Global Multiregional Input-Output Frameworks: An Introduction and Outlook. *Economic Systems Research* 25(1): 1–19. http://dx.doi.org/10.1080/09535314.2012.761179.
UN. 2008. *System of National Accounts 2008*. New York.
Weidema, B.P. 2009. New options for material flow accounting in the ecoinvent database. https://www.ecoinvent.org/files/200909_weidema_new_options_for_material_flow_accounting.pdf. Accessed June 21, 2018.
Wernet, G., C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz, and B. Weidema. 2016. The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment* 21(9): 1218–1230. https://doi.org/10.1007/s11367-016-1087-8.