Freiburg, Germany, February 2019

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

Glossary and FAQ

Version 19-Feb-2019

Stefan Pauliuk, Niko Heeren, Mahadi Hasan, and Daniel B Müller

Glossary

Aspect: To locate data in a system definition, one has to specify the aspects of the data that describe how the given values relate to time, region, processes, etc. in the system. The aspects describe how the system dimensions relate to the data. For example, a flow has a starting node and a terminating node. That means that 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 stocks in a system. [From paper]

Aspect domain: An aspect domain is a set containing the different values that a certain dataset aspect can take. For example, a dataset on emissions of fossil CO_2 to air by region and time has the aspect 'time', and the set of all years that the time aspect can take, e.g., [1950, 1951, … 2016, 2017], is the aspect domain of the time aspect for that particular dataset. The term domain was chosen because it is used in mathematics to define functions (mappings).

Classification: A breakdown of a system dimension into discrete units, e.g., the breakdown of time into different years.

Classification item: A member of a classification, e.g., the ISO region 4 (Afghanistan) in the ISO 3166 region classification.

Data point: Individual value (+ uncertainty, unit) quantifying a fact in a system.

Dataset: Collection of data points with the same aspects and source. E.g., "product lifetimes from xxx study."

Data group: A collection of datasets from a common source, figure, table, or research project.

Data model: Each data type (stock, flow, material content, product lifetime, …) has a specific data model that prescribes which aspects are required and which aspects are optional for the meaningful location of this data type in the system definition. [From paper]

Data category: One of the six broad data groups defined in the paper: stocks, flows, intensive object properties, intensive process properties, extensive process properties, and general ratios.

Data type: Specific class of information under a data category, e.g., flows, process inventories, or unit process inventories as part of the 'flow' category, or product lifetime and material composition as part of the 'intensive object properties' category.

Dimension: "A dimension is a structure that categorizes facts and measures...dimensions provide structured labeling information to otherwise unordered numeric measures. The dimension is a data set composed of individual, non-overlapping data elements. The primary functions of dimensions are threefold: to provide filtering, grouping and labelling." From https://en.wikipedia.org/wiki/Dimension_(data_warehouse). In the general data model, the term dimension is used to categorize the measures and facts that are needed to locate data in a system definition, such as time, region, process or material.

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 or to group events, the object 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. [From paper]

Examples and explanations of the propositions 1-5 for the general data model of socioeconomic metabolism

(P1) The data point with the numerical value “1620.4 Mt/yr” is meaningless until it is specified how it relates to the different system dimensions of socioeconomic metabolism.

To assign a meaning to this data point, one must say that the regional scope of the data point (or is region aspect) is ‘global’, the time period aspect is 2015, the material aspect is ‘crude steel, the process of origin is ‘crude steel production’, and the process of destination is ‘steel forming’. With these aspects specified, it becomes clear that the data point describes a flow of crude steel from its production to its use, commonly labelled ‘crude steel production’.

Cf. Table S19 for a complete list of mandatory and optional aspects of data.

(P2) Starting from the example above, we generalize the phenomenon described by the production volume of crude steel and define a data type ‘flow’ with the following core aspects and their meaning:

[object]: material, substance, product, or commodity that flows.

[origin_process]: process where flow originates from

[destination_process]: process where flow terminates

[time]: time period over which flow is measured

[region]: geographical scope over which the flow of objects is measured/quantified

(P3) Staying with the flow example, we define the set containing one material (A_1 = {‘steel’}), the set of the processes {‘crude steel production’, ‘steel forming’} in the steel industry (A_2), the set of all years between 1900 and 2018 (A_3), and the set of all countries (A_4).

Then, we can define the dataset WHCSP (‘world historic crude steel production’, data type = ‘flow’) as a function/mapping as follows:

$$WHCSP : A_1 \times A_2 \times A_2 \times A_3 \times A_4 \rightarrow \mathbb{R} \cup \{\text{null}\} \quad (1)$$

Where the list of aspect domains and their meaning is given by the definition of the data type ‘flow’ above. Note that the domain A_2 appears twice, as the system dimension ‘process’ appears twice in the data type flow: as origin_process and destination_process.

The individual data points in this dataset are written as tuples that the function WHCSP maps into the real numbers plus ‘null’.

WHCSP(‘steel’, ’crude steel production’, ’steel forming’, ’global’, ’2015’) = 1620.4 Mt/yr

WHCSP(‘steel’, ’crude steel production’, ’steel forming’, ’India’, ’2017’) = 101.4 Mt/yr

[…]

Tuples that are assigned no value point to ‘null’, which can have several reasons:

WHCSP(‘steel’, ’crude steel production’, ’steel forming’, ’Brazil’, ’1915’) = ‘null’ # no data in this dataset

WHCSP(‘steel’, ’crude steel production’, ’steel forming’, ’Brazil’, ’1315’) = ‘null’ # no data, neither Brazil nor crude steel production existed in 1315. Tuple not meaningful for this particular dataset.

WHCSP(‘steel’, ’steel forming’, ’crude steel production’, ’global’, ’2015’) = ‘null’ # other flow direction. Tuple not meaningful for this dataset.

(P4) The set of all countries (A_4) defined above is a classification of the world into different countries. Depending on which regional classification is chosen (ISO country list, UN world regions, etc.) the data will be different as they would be reported for different aggregation levels. That means when handling datasets not only their data model but also the classifications used must be specified.

(P5) To make data interoperable across methods and publications their respective data types should be clearly identified. Often it is convenient to group data of different types together, e.g., in figures, tables, MFA systems, or data files (like the ecospold file format for activity data). While the core data model keeps data of different types separate, the data group allows for the bundling of data of different types.

Additional explanations and frequently asked questions

Why do we need this? Doesn't LCA already have a comprehensive data model in form of the ecospold format?

Doubtlessly, ecospold is an elaborate and comprehensive data format. It not only contains process inventories (exchanges) but also price and production volume of reference flows and the composition (carbon content, water content) of substances. But there are many more data that are needed in industrial ecology research, including: product lifetimes, breakdown of products into components and materials, process capacities, population, economic statistics, trade flows, or economy-wide and sector-specific material flow accounts. For those data there is currently no general data model, and this gap is filled by the data model presented here.

Why do we need to distinguish between system dimensions and aspects?

The first reason is that it is an important insight for the data modeller to understand the difference between a system dimension (general category or measure) and the different roles they have when locating data in the system. The second reason is that classifications (e.g. ISO code for regions) always apply to the system dimensions, and by linking data aspects to dimensions, the data modellers can then use the classifications of the dimensions to describe the different aspects.

How does the general notion of systems relate to the system definition (boxes and arrows) commonly used in industrial ecology research?

As a complex system, any conceptualization of SEM contains the following basic system elements (von Bertalanffy 1968; Forrester 1968): i) nodes, ii) links or relations between nodes, and iii) the system boundary. Nodes (or vertices) linked by edges (or arrows) represent directed graphs (Diestel 2012), and this type of graph is the basic structure of system definitions of SEM used in IE research (Pauliuk et al. 2015). Nodes represent industries (transformation processes), markets (distribution nodes), or storage processes (nodes with stocks). Edges represent flows of goods, products, or substances between nodes and across the system boundary (Pauliuk et al. 2016).

How do I get data into the database?

By filling them into the excel template (use one that you downloaded and fill with new data and their description) and email them to in4mation~@~indecol.uni-freiburg.de. We'll review them and respond depending on our workload. After the data have passed review, they are uploaded with a Python parser available on the GitHub repo. It is also possible to write custom upload scripts for datasets where reformatting as Excel template is not practical. Make sure that the license allows for redistribution. Cf. also the section "Steps of contributing data to the IEDC" below.

Do the data need to come in a standard classification for regions, products, etc?

No. The guiding design principle of the IEDC is to be prescriptive about the data structure and types, but flexible about the resolution. This way, we can cast the available information into a structure that is commonly understood, but still keep the flexibility to use the resolution needed for the research question at hand.

We do use common classifications such as region ISO-codes or NACE commodity groups, and encourage data providers to use those wherever convenient. If data are provided in a standard classification, they will automatically be matched. If not, a custom classification will be created.

Aren’t the system location data also metadata?

While it is clear that system location data are very different from information on data description, provenance, source document, author and version information, or license, it was not obvious to us whether these data should also be considered metadata or not. System location data fit the definition of metadata: "Metadata is data [information] that provides information about other data". [https://www.merriam-webster.com/dictionary/metadata] But system location data is not one of the 'types of metadata' that are commonly listed as examples. A strong argument against defining system location data as metadata follows from the definition of 'system location' used in the paper:

> “ 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, …)"

This 'system location' data is basically the aspect tuple which is needed to give meaning to any value in a system, which, in our opinion, resembles data more than metadata. Hence, we decided to not label system location data as metadata.

How does the database model work?

A common relational database model for tuple-based data and OLAP cubes is the star schema database (Adamson 2010). Here, each data aspect in the data table that contains the OLAP cube is linked to a dimensional table, which describes the different aspect items, e.g., by listing the population and area of the different ISO regions. The star schema database structure served as template when developing the IEDC relational database model, but some modifications were necessary.

First, 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 of the different data table entries. The dataset table contains a complete description of the dataset, covering metadata, system allocation, a dataset description, and an ID. The data table has the same features, but for individual data items within the dataset. Datasets can be part of data groups (e.g., different flow dataset together describe an MFA system, or flow data, price and production volumes, and water content data together form a unit process inventory), and data groups can be part of projects (e.g., for each paper of a cumulative dissertation (project), there is one data group containing the different datasets).

Second, while the data items point 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 further link to dimensions, and the data types further link to the data categories 1-6. Lookup tables without further links are used to describe data sources, provenance, users, uncertainty, etc.

Why a relational database?

A relational database is just one way of organizing the data; it was chosen because of the ease of designing the database and for the convenience of the MySQL tool chain. SQL is a powerful language that allows us to use different datatypes, set automatic increments for IDs, create uniqueness constraints for columns, or create links to ids of other tables. Different formats such as graph databases and Resource Description Framework (RDF) databases (‘triple stores’) are possible, but require more exploration and definitely more resources to implement. The entire content of the project is open so everybody is welcome to experiment with other database models that implement the data model for socioeconomic metabolism and chip in their experience!

What do the different data aspects mean?

Data aspects link data to system dimensions. Data aspects answer questions like ‘what’ (what is flowing or what is emitted), ‘where’ (where does material come from?), ‘when’ (when was the stock measured, of which age-cohort were the products recorded?), and ‘how’ (in what unit/layer is the data measures, how does the data relate to other variables (like input per output flow x))?

The screenshot from the aspects table of the IEDC, taken from http://www.database.industrialecology.uni-freiburg.de/aspects.aspx

on Jan 28, 2019, shows the hitherto defined aspects and their meaning (Fig. S1):

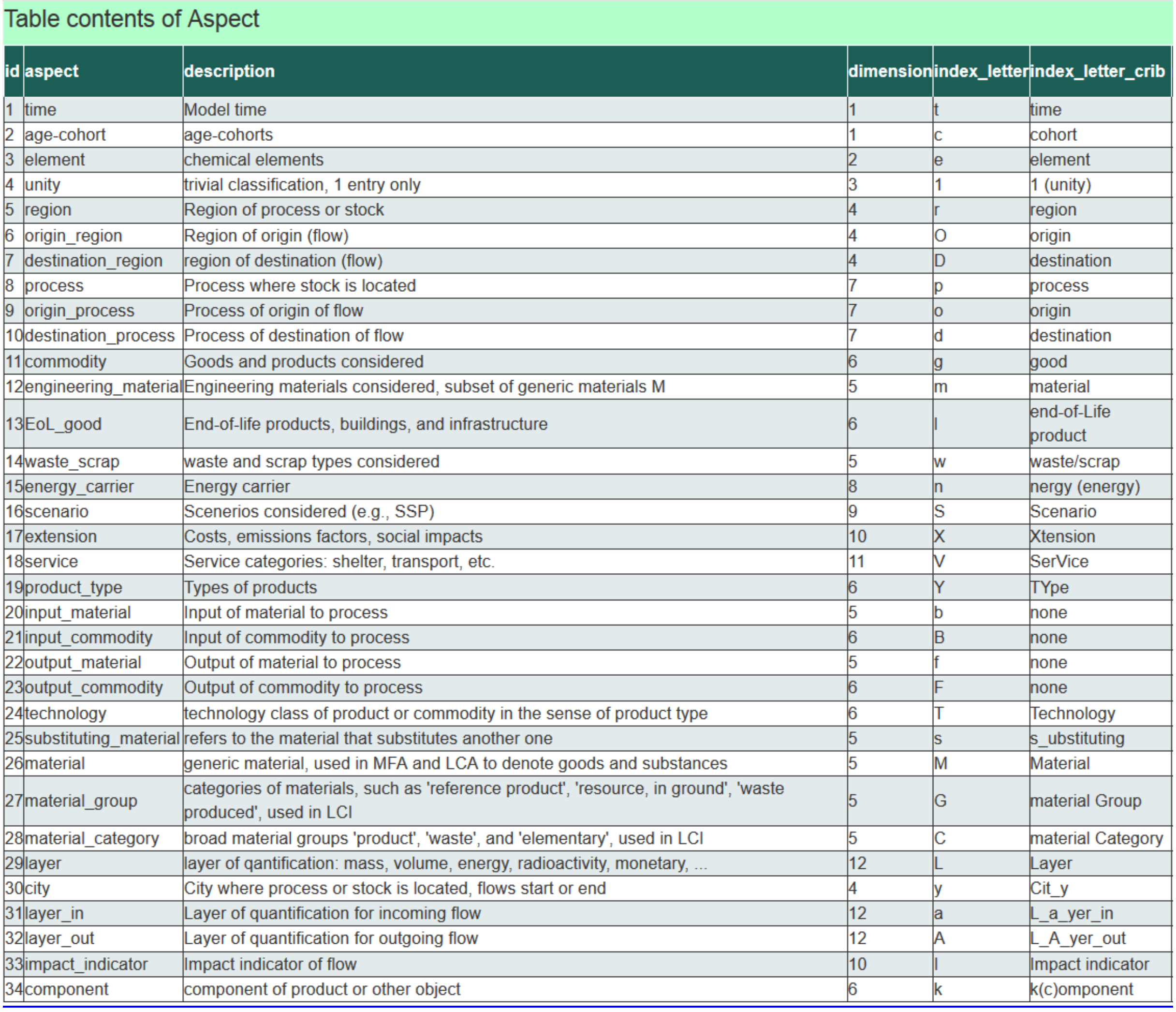

Table contents of Aspect

id	aspect	description	dimension	index_letter	index_letter_crib
1	time	Model time	1	t	time
2	age-cohort	age-cohorts	1	c	cohort
3	element	chemical elements	2	e	element
4	unity	trivial classification, 1 entry only	3	1	1 (unity)
5	region	Region of process or stock	4	r	region
6	origin_region	Region of origin (flow)	4	O	origin
7	destination_region	region of destination (flow)	4	D	destination
8	process	Process where stock is located	7	p	process
9	origin_process	Process of origin of flow	7	o	origin
10	destination_process	Process of destination of flow	7	d	destination
11	commodity	Goods and products considered	6	g	good
12	engineering_material	Engineering materials considered, subset of generic materials M	5	m	material
13	EoL_good	End-of-life products, buildings, and infrastructure	6	l	end-of-Life product
14	waste_scrap	waste and scrap types considered	5	w	waste/scrap
15	energy_carrier	Energy carrier	8	n	nergy (energy)
16	scenario	Scenerios considered (e.g., SSP)	9	S	Scenario
17	extension	Costs, emissions factors, social impacts	10	X	Xtension
18	service	Service categories: shelter, transport, etc.	11	V	SerVice
19	product_type	Types of products	6	Y	TYpe
20	input_material	Input of material to process	5	b	none
21	input_commodity	Input of commodity to process	6	B	none
22	output_material	Output of material to process	5	f	none
23	output_commodity	Output of commodity to process	6	F	none
24	technology	technology class of product or commodity in the sense of product type	6	T	Technology
25	substituting_material	refers to the material that substitutes another one	5	s	s_ubstituting
26	material	generic material, used in MFA and LCA to denote goods and substances	5	M	Material
27	material_group	categories of materials, such as 'reference product', 'resource, in ground', 'waste produced', used in LCI	5	G	material Group
28	material_category	broad material groups 'product', 'waste', and 'elementary', used in LCI	5	C	material Category
29	layer	layer of qantification: mass, volume, energy, radioactivity, monetary, ...	12	L	Layer
30	city	City where process or stock is located, flows start or end	4	y	Cit_y
31	layer_in	Layer of quantification for incoming flow	12	a	L_a_yer_in
32	layer_out	Layer of quantification for outgoing flow	12	A	L_A_yer_out
33	impact_indicator	Impact indicator of flow	10	I	Impact indicator
34	component	component of product or other object	6	k	k(c)omponent

Figure S1: IEDC aspect table, Jan 28, 2019. Each aspect is assigned a unique letter code, that can be used in variable notations, like denoting a certain flow as 1_F(g,o,O,d,D,t): Flow 1_F of commodity g from origin_process o and origin_region O to destination_process d and destination_region D in time period t.

Does the database engine check whether the aspects for a given dataset are correct and complete?

No. At the moment, there is no routine to check whether all mandatory aspects for the different datasets are provided. There is also flexibility in the aspect description to accommodate for the large diversity of data source in our field. It is the responsibility of the authors to build meaningful and complete data models (in form of the aspect structure) of the data they submit to a database, and it is the responsibility of the data reviewers to check the correctness and completeness of the data model for each dataset submitted, and to request changes if the model is ambiguous or incomplete.

Why a standalone database that is not linked to the semantic web or other classification standards?

Many IE datasets come in custom classifications that are not linked to established industry, product, material, or other classifications. Hence a stand-alone data description is necessary to reformat and store these data in a common database.

For those datasets that use established classifications or common items (like NACE, years, or chemical elements) a standard classification should always be used. Those standard classifications are entered before uploading the datasets, and they contain standardized reference keys (id) such as NACE codes or chemical element symbols. It is also possible to insert a Uniform Resource Identifier (URI) as one of the classification item's attributes, and thus establish a link to semantic webs.

Summarizing: You don't have to link to established classifications or semantic webs for datasets that come without such a link for some of their aspects, but you can easily establish such a link by a) using standardized classifications and b) adding URIs as classification item attributes for the items of those standard classifications.

Why can the data layer (mass, volume, energy, items, …) be both: a dataset attribute and an aspect?

That is a convenience feature. In some fields, quantification on one layer only is the standard, like in MFA. Therefore, by default, each dataset needs to have a global 'layer' attribute. There is then no additional layer aspect in the data model for such datasets.

Other datasets contain data on multiple layers, like unit process inventories (energy, water volume, mass, items, radioactivity, …), and to be able to enter those conveniently, the global 'layer' attribute is set to 'Misc. units' or 'Misc. physical units' and the actual layers are then given in the 'layer' aspect of the dataset.

What is the difference between the different flow data types: 1_F, 1_PI, 4_UPI?

What is the difference between and exchange and a flow?

There are some subtle differences between those datatypes, which all roughly are described as 'flows':

Flows in MFA or SNA (System of National Accounting) are different from LCA exchanges: The former ones describe events of material/products moving from one process to another during a given time interval. The latter describe an amount of input or output relative to a reference output or input (unit process). These data are constructed from actual flow measurements but are reported as being valid for a certain geographical scope and time frame, and don't have to be connected to a process or origin/destination.

For 4_UPI the ‘time’ aspect refers to the time period for which the dataset is considered an accurate description of the real situation, but for 1_F and 1_PI data, it refers to the actual measurement interval. Note that unit process inventories (UPI) are part of data category 4, as they are normalized.

Unlike for ‘basic’ flows of type 1_F, for 4_UPI and 1_PI datasets one also needs to specify two LCA-specific aspects: material_category (elementary or intermediate) and material_group (From or to environment, reference product, waste, etc.).

Moreover, the entries in 4_UPI and 1_PI datasets can have an undefined origin (for inflows to process) or undefined destination (for outflows from process), which is not possible for 1_F flow data.

Flows (1_F_) describe actual events (objects flowing from one process to another during a time interval), whereas exchanges (1_PI_ and 4_UPI_) describe relative inputs and outputs to/from a process/activity, where the origin/destination can be unknown and the time interval does not have to be the interval of observation but denotes the time period for which the data are valid.

What is the difference between the different price data types 3_PR and 6_FPR?

Prices (and other intensive quantities) can be recorded in two ways: a) as intensive object property, e.g., the price tag of a product in the shelf of a supermarket: 3_PR_ (product/material, time, region) b) as intensive flow property: e.g., the price of cement imported from country A into Country B. This dataset is the ratio of two layers of the same flow: monetary / mass or items. It belongs to the general ratio category 6: 6_FPR_ as it represents the ratio of two flow layer measurements. Recorded with the same aspects as a flow, just a different layer. Prices are always extrinsic, they measure an effect of the system (i.e., demand-supply relations) that is allocated to a good at a specific time and place. A bottle of water might be worth much more in the desert than in an urban area with ample supply – it’s the systems context that determines the price.

What is the difference between the different data types for recording environmental extensions: 1_F, 1_PI, 4_UPI, and 4_PE?

‘Environmental extensions’ refer to information about resource use and emissions to the environment of industrial or other processes in the technosphere/anthroposphere. These data come in different forms, some of which have subtle differences which means they are stored as different datatypes in the IEDC. In particular:

+ Measured or reported flows of emissions and resources, like domestic material extraction in region x in year y or the CO2 emissions of power station x in year y are recorded as flows 1_F.

+ If the extensions are part of a process inventory also recording technosphere inputs or waste, these data together are recorded as process inventory 1_PI, which contains flow data on different layers of measurement (mass, energy, items), which have in common that they all enter or leave the same process/activity.

+ If the extensions as part of a process inventory dataset are normalized for one reference product these data are recorded as unit process inventory 4_UPI. Here the unit is still the unit of a flow but the meaning of the data is that it is flow quantity per given quantity if the reference output. Moreover, the time aspect of the UPI is NOT the measurement time interval but the interval for which the UPI data are representative.

+ Extensions also come as coefficients and not as UPI, as typical in the engineering literature, like CO2 per ton of pig iron in blast furnaces, iron or per ton of sintered ore, or water per ton of crop harvested. Unlike UPI data, these data then have the unit of coefficients, like ton/ton, which is why a dedicated data type 4_PE was created. By dividing and extension stored in a UPI by the reference flow of that UPI, a 4_PE coefficient is created.

The flexibility in datatypes is needed to accommodate for the specific formats that have evolved in the different fields, and to distinguish between measurements and coefficients.

How can existing data formats, like ecospold, MRIO tables, or STAN data, be mapped to the IEDC data model and structure?

We compiled a mapping of ecospoldv2 and the LCI data ontology by Kuczenski et al. (2016) to the IEDC structure, cf.

https://github.com/IndEcol/IE_data_commons/blob/master/doc/LCI_IEDC_Correspondences_V1.xlsx

For MRIO, we have specific examples in the prototype, just search for 'EXIOBASE' or 'IOT', and a mapping of the different MRIO tables in the mapping file linked above.

STAN files (.mfa) are zipped xml files that contain information in plain text (time and regional scope of system, flow values and which processes they are linked to) and encapsulated binary data (containing the arrangement of flows and processes on the system canvas). Thus it is possible to parse STAN xml files to extract flow data and format them to the IEDC format.

Can a process (like steel production) be quantified in two different regions at once? Is there not a danger for (inconsistent) data redundancy, and incompatibilities with ecospold2, in having processes as dimensions rather than datatypes?

Here, we need to distinguish between an "activity dataset"/"activity description", which does have a regional and temporal scope, and the activity name, which in the IEDC language is a process, without region and time.

Consider the following random example from the ecoinvent 3.3. dataset

000b493f-5e76-4fe0-93ca-73b49263c1fc.spold

This activity dataset has the following tags/attributes:

+ <activityName xml:lang="en">chlorine dioxide production</activityName>

-> This is a process in the IEDC.

+ <geography geographyId="34dbbff8-88ce-11de-ad60-0019e336be3a"> <shortname xml:lang="en">GLO</shortname> </geography>

-> This is the geographical/regional scope of the dataset, not the process itself! All IEDC datasets extracted from this .spold file will have this regional scope.

+ <timePeriod isDataValidForEntirePeriod="true" endDate="2000-12-31" startDate="2000-01-01"/>

-> This is the temporal scope of the dataset, not the process itself. All IEDC datasets extracted from this .spold file will have this temporal scope.

It is not correct to say, for example, that "an activity occurring in Canada would be represented by a different process than a similar activity occurring in Germany". The _datasets_ describing these similar activities would have the same _process aspect_ in both cases (e.g., chlorine dioxide production), but different regional scope. The process attribute in ecospold does not 'own' the geography information, as can be clearly seen from the ecospoldv2 specification. Instead, the activity dataset owns the geography information! The IEDC datasets extracted are located in space and time (they have regional and temporal aspects) that correspond to those of the activity dataset.

Therefore, a process (like chlorine dioxide production or car driving) can be in two different regions at once, and the datasets describing these parallel activities all have region aspects to capture regional diversity.

By separating data into region, time, and process aspects redundancy is reduced, not increased, as each process, region, time needs to be defined only once and these definitions are then combined to meaningfully describe datasets, like production flows of process type A in country R vs. production flows of process type A in country S. Note that this is not a new IEDC feature, ecospold has it as well, see my example of the ecoinvent specification above.

The difference between ecospoldv2 and IEDC datatypes is that the former groups different datasets (process inventory, prices, production volumes, impact indicators, …) into a single data file (data group in IEDC), so that the process, region, and time aspects apply to all data in there and thus are specified only once. In contrast, the up to seven IEDC datasets that can be extracted from ecospoldv2 each need to specify the process, region, and time aspects. The organisation of data is different: ecospoldv2 has seven IEDC data types for one single (process/region/time) aspect tuple, and the IEDC has seven data types for many (process/region/time) aspect tuples.

Why is there no command or filter to identify LCA, MFA, or IO data?

A main reason for having a database such as the IEDC is to encourage researchers to stop thinking of data as belonging to methods, but as describing the underlying system. Cf. first part of paper 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 events 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."

That is why there is no 'method' attribute in the IEDC, and method-specific databases already exist.

Of course, data providers can supply keywords or comments for their data, indicating previous or typical uses or methods that these data come from. In our opinion, this option is enough to help researchers to quickly find data that serve their needs.

How are new data types, aspects, classifications, etc., created?

When a dataset is prepared for upload, it can be that it does not well fit to the existing data types, the aspects, dimensions, and layers needed to describe it have not been defined yet, or the classification used has not been uploaded yet.

In this case a dialogue with the database managers is needed to find out whether the meaning of existing dimensions, aspects, data types, and layers can cover the incoming data (preferred option) or whether new lookup table items need to be created indeed. It can also be found that certain data should not be part of the IEDC but organized in other databases, for example, bibliographical information.

Why do materials, commodities (products), energy, and services constitute separate dimensions and are not aspects of a more general ‘object’, ‘flowable’, or ‘stuff’ dimension?

Both options would work. There were longer reflections and discussions on whether a general ‘object’/’stuff’/’flowable’ system dimension should be used, which comprises everything that can flow, including substances, (engineering) materials, products, energy, or services. This choice would be the preferable option from an abstract system modelling point of view.

In practice, however, one would have many data aspects linking to the same dimension, for example, a dataset describing the energy requirements of turning ethylene into polyethylene would link to the ‘stuff’ dimension three times. Instead, we believe that having explicit dimensions for materials (when the physical properties are most relevant), commodities (when the fact that stuff is traded and used is most important), energy (when the energy-carrier property is most important), and services (when the non-material service property is most relevant) makes the data model more tangible and useful.

We will now gain experience with the detailed option and review our choice at a later stage.

What are the different “attributes” of a classification item? Why do they have different datatypes?

Each classification item (think ‘China’, ‘Aluminium’, or ‘electric vehicle’) can have up to 15 different attributes. For example, Aluminium has a symbol (Al), an atomic number (13), both of which are unique, and other attributes, such as atomic weight. A country has a time period during which is exist(ed), several numeric and letter codes, an area, etc. The attributes ending on ‘oto’ (one-to-one) are reserved for unique and thus defining attributes, e.g., for names, symbols, and atomic numbers of chemical elements. These can be referred to by the incoming datasets, meaning that one dataset submitted for upload can have region ISO codes for the different regional aspects of the data it contains, and another one can have (standard) region names. The database upload parser is able to handle both and any other labels supplied as _oto attributes. The ‘anc’ (ancillary) attributes contain additional information. Uniform Resource Identifies (URI) can be entered as oto attributes. There must be at lead one unique oto attribute for each classification to label the item, which is why the pair ‘attribute_1_oto’ and ‘classification_id’ must be unique. The default data type is a string of up to 255 bytes, but for attribute_2_oto we switched to text to accommodate the often long category names in the established commodity and industry classifications.

Where exactly is the boundary between the general data model and its flexible practical implementation?

The data model states that each data type (stock, flow, coefficient, price, capacity, …) has a set of aspects (time, region, age-cohort, material, …) that link to the dimensions of the system where the data a located.

From the need to describe objects (products, materials, commodities) and processes (transformation, distribution, storage, use) a set of seven general data categories (cf. Table 2 of the paper or http://www.database.industrialecology.uni-freiburg.de/datatypes.aspx) was defined.

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 (stock, flow, process capacities, yield coefficients, impact indicators, …) the result of consensus-building among data providers. The current IEDC application contains 28 data types, which have been identified while investigating the data commonly used in the different IE subfields.

Why are correspondence tables considered data and added as separate data category 7?

Adding correspondence tables to the IEDC enables us to store this important type of information and prepare the database for automatic matching between datasets using different but corresponding classification. The question was then whether correspondence tables should be stored in separate correspondence_definition and correspondence_items tables, or whether the existing datasets and data tables should be used. The former option would be preferable from a data modelling point of view as it keeps data and their aspect classifications separate. The latter is easier to implement and does not lead to the creation of new tables, thus reducing the database structure to the core functions needed, which is also why we – for the prototype – decided to simply implement a new data category.

Why is the lifetime of a product/material/substance an intensive object property and not an intensive property of the process of residence?

Once could describe the lifetime as an intensive property of processes, not the objects. The lifetime is a statistical relationship between inflows and outflows of a process. There are many aspects that determine the lifetime of a building, for example: the building itself, but also its use, how it is maintained, regulations, other interests in the land it stands on, etc.

We allocate a lifetime to a product when we do the modelling, but it is in reality not the product that determines its lifetime.

In the IEDC, it is exactly this allocation of lifetime to products that we want to describe. Since [process] is a mandatory aspect for the lifetime, the aspect structure [material/commodity] in [process] in [region] works well for the actual data model where lifetime is an intensive (per unit) property of an object in a process.

Why is the term 'object' used and what exactly is it? For example, is carbon an object and if yes, what do we mean by that?

We need the 'object' dimension to refer to all the 'stuff' that moves around in the system studied. We understand 'objects' as substances, materials, products, goods, or commodities, depending on the context in which they are described (e.g.: CO_2 in the atmosphere: substance, CO_2 in cartridges for making sparkling water: commodity).

For example, is carbon an object and if yes, what do we mean by that? Do we mean (i) one carbon atom, (ii) all carbon atoms (including the ones on other planets), or (iii) the carbon embodied in a specific good? In the last case, is it not an aspect of a specific good rather than an object?

Yes, carbon (the chemical element) is an object in our data model. The question is now: How do we link carbon to a system studied and to the data describing that system? There are three possibilities. First, carbon is a classification item of the 'element' system dimension, and we would use that version of carbon as data aspect whenever we refer to the chemical element carbon in the system, e.g. when indicating the carbon content of a commodity in a flow, that is option (iii) above.

Second, we can also specify carbon as a classification item of the material dimension and use that as aspect when singling out one or more carbon atoms in flows or stocks (option (i) and (ii) above).

Third, we can specify carbon as classification item of the commodity dimension when referring to commercial flows of carbon. In these cases, however, it is often better to use the more descriptive commodity groups such as 'graphite' or 'black carbon' instead.

In the IEDC, carbon can be a chemical element, material, or commodity depending on how it appears in the data models used.

Which conservation laws hold for the different objects?

Baccini and Brunner (1991) apply the law of mass and energy conservation to industrial systems analysis. In practice, that entails a conservation law for each individual chemical element for all processes without nuclear reactions. For certain processes, one can also apply a “law of object conservation”. For example, alloys and other embedded materials in products often do not change their composition during product use, even during some waste management processes. Manufactured goods stay largely the same during transport and storage, some even during the use phase. From a closer perspective, however, the consumption of new goods in a given time period is different from the outflow of used goods (in size, cohorts, composition…), as the goods may have changed significantly during the use phase (e.g., exchange of parts for maintenance, some parts may have rusted, or others may be missing, and also the price is likely to change during the useful lifetime).

Hence the use of ‘higher-level conservation laws’ for materials and products can be practical and convenient for some applications, but can also be very limiting for others. We therefore see Baccini’s definition as more basic and valid for a broader range of applications, but offer the flexibility in the IEDC to trace different objects and apply conservation laws where convenient, e.g. by allowing flows and stocks to be defined for both: materials and products.