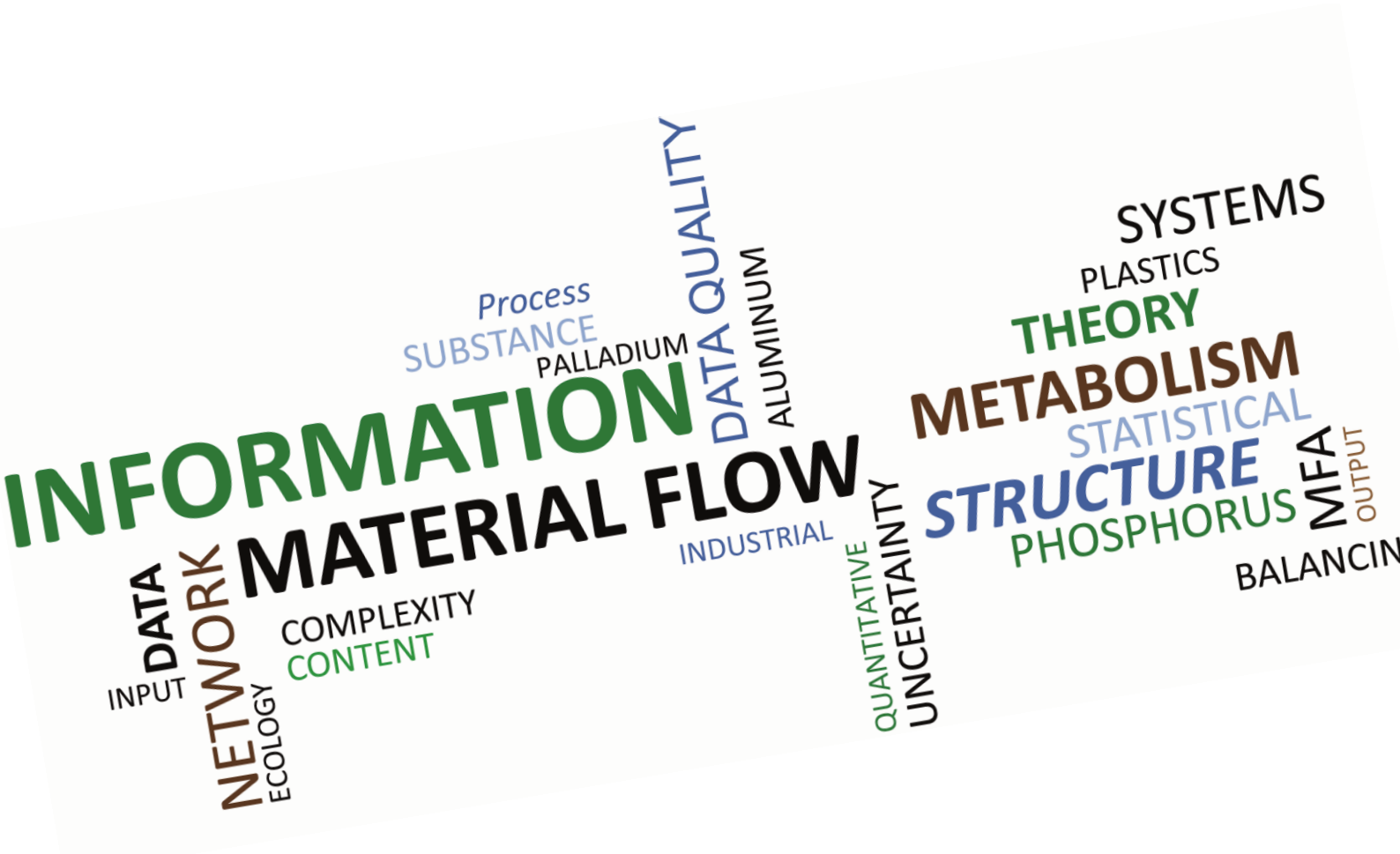
Systematic Evaluation of Data, System Structure and Information Content in Material Flow Analysis

Dissertation of Oliver Schwab
Technische Universität Wien
Vienna, 2016



Corrigendum to the dissertation

"Systematic Evaluation of Data, System Structure and Information Content in Material Flow Analysis"

submitted by Oliver Schwab at TU Wien, Vienna, 2016

The author discovered five flaws, which are listed and corrected in the following.

1. In Eq. 11 (page 39), the indexation of the sums is inconsistent. Eq. 11 should be, with $i=\{1,\ldots, n_F\}$,

$$S = -\sum_{i=1}^{n_F} Fi \cdot \sum_{i=1}^{n_F} \frac{Fi}{\sum_{i=1}^{n_F} Fi} \log \frac{Fi}{\sum_{i=1}^{n_F} Fi} = -n_F \, log \frac{1}{n_F}$$

2. There is a typo in Eq. 12 (page 41), where it says that the denominator of the fractions is $\sum_i Fi$. This should be $\sum_{i=1}^{n_F} ID_{Fi}$. (1) holds also for Eq. 12, which should be

$$U_{ap} = -\sum_{i=1}^{n_F} ID_{Fi} \cdot \sum_{i=1}^{n_F} \frac{ID_{Fi}}{\sum_{i=1}^{n_F} ID_{Fi}} \, log \frac{ID_{Fi}}{\sum_{i=1}^{n_F} ID_{Fi}} = -\sum_{i=1}^{n_F} ID_{Fi} \, log \frac{ID_{Fi}}{\sum_{i=1}^{n_F} ID_{Fi}}$$

3. There is a typo in Eq. 13 (page 42). Instead of the logarithm of $\frac{ID_{Fi,b}}{\sum_i X_{Fi,b}}$, it should be the logarithm of $\frac{ID_{Fi}}{\sum_{i=1}^{n_F} ID_{Fi}}$. (1) also holds for Eq. 13, which should be

$$U_{b,w} = -\sum_{i=1}^{n_F} \frac{X_{Fi,b} n_F}{\sum_{i=1}^{n_F} X_{Fi,b}} ID_{Fi,b} \, log \frac{ID_{Fi,b}}{\sum_{i=1}^{n_F} ID_{Fi,b}}$$

4. Remark to Table 6 (page 44) and the associated text: There is one more complex group of topologies than the one illustrated in example D. In this group of topologies, every process connects also to itself. Such a topology with n_P processes has $n_{F,max}=n_P^2$ flows and it is $C=S$ and $T=0$.

5. In Figure 20 (page 51), the data attribute "producer type" is plotted. Correctly, as it says in the caption, the data attribute "origination category" should be plotted, as illustrated below.

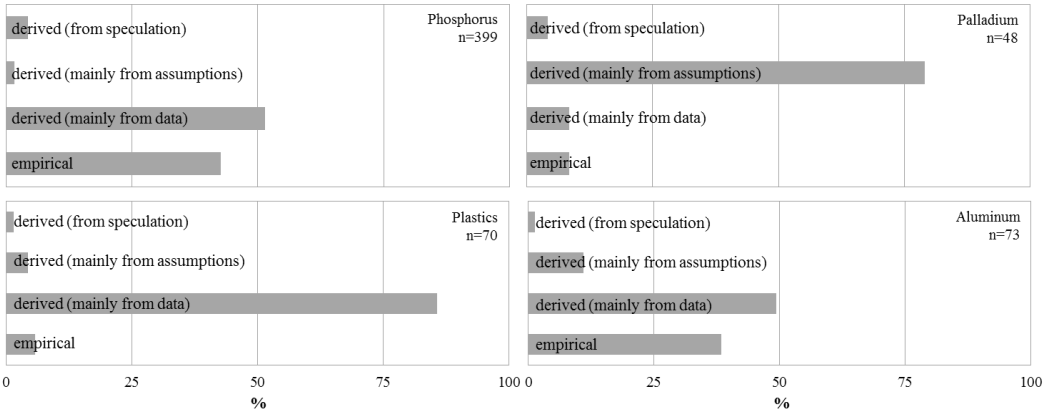


Figure 20: Origination category of data elements.

Doctoral Thesis

Systematic Evaluation of Data, System Structure and Information Content in Material Flow Analysis

submitted in satisfaction of the requirements for the degree of
Doctor of Sciences (Doctor technicae)
of the Vienna University of Technology, Faculty of Civil Engineering

---

Dissertation

Systematische Bewertung von Daten, Systemstruktur und Informationsgehalt in der Stoffflussanalyse

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften (Doctor technicae)
eingereicht an der Technischen Universität Wien, Fakultät für Bauingenieurwesen
von

Dipl. Geoökol. Oliver Schwab
Matrikelnummer 1328710
1040 Vienna, Austria

Scientific committee

Prof. Helmut Rechberger
Institute for Water Quality, Resource and Waste Management
Technische Universität Wien, Austria

Associate Prof. Reinout Heijungs
Department of Econometrics and Operations Research
Vrije Universiteit Amsterdam, The Netherlands

Associate Prof. Gang Liu
Institute of Chemical Engineering, Biotechnology and Environmental Technology
Syddansk Universitet, Denmark

Vienna, December 2016

To know one's ignorance is the best part of knowledge.

Laotse

Abstract

Material Flow Analysis (MFA) is a useful method for modeling, understanding and optimizing material flow systems. MFAs incorporate databases of increasing size and quality and reveal more and more details about material flows into, within and out of given systems. As a consequence, MFAs are of increasing size and system structures are of increasing complexity. Due to differences in data quality, it is not always clear how reliable MFA results are.

In this thesis, uncertainty and complexity in MFA are approached from a system-theoretical perspective and formalized as measures for characterizing and distinguishing material flow systems by their information content and system structure. MFAs are, in a graph-theoretical sense, understood as networks. The information content and system structure of these networks are described by formally linked metrics derived from the field of theoretical ecology. The structure of a system is computed according to the configuration of each individual flow in relation to its neighboring flows. Integrating measures for data quality, the uncertainty of quantitative MFAs before and after balancing is determined and the information content of material flow systems is quantified. As the applicability of statistical measures for the evaluation of data quality is typically limited in MFA, it is proposed to approximate data quality by means of multi-dimensional functions of MFA data attributes. Data attributes are data-associated annotations concerning statistical properties, meaning, origination and application of the data. These data attributes are systematically documented and evaluated in a data characterization matrix, which forms the basis for automated estimation of data quality and subsequent quantification of information content.

Exemplarily, four material flow systems (phosphorus, palladium, plastics and aluminum) are analyzed, compared and distinguished in terms of their information content and system structure. The proposed procedures are useful for gauging the information content of MFAs and for analyzing their system structure by means of quantitative measures. They contribute to a better understanding of the informational basis of material flow systems. They enable material flow systems to be compared to one another and changes in the information content of material flow systems over time to be tracked. The proposed measures support the design of MFA systems, optimized use of available information, communication of MFA results, and decision making in scientific and institutional contexts in light of limited information.

Preface and author's contribution

Soon after I became part of the team at TU Wien in 2013, Professor Helmut Rechberger raised a question that should not leave me alone for more than the following three years: "It would be interesting to know the information content of MFAs". He annotated that this may be a rather abstract research object. While intuitively agreeing on this annotation, I started framing "information" as an object of research within the general context of MFA. Thankfully, the Austrian Federal Ministry of Science, Research and Economy provided funding for this attempt in the course of a project series called EDNA (Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen (Investigation of the data requirements of national resource budgets)).

Before long, I noticed that some MFAs are based on extensive, credible databases and others lack reliable model input information. This initial difference seemed to get lost during preparation of the MFAs and did not always reflect in MFA results. A measure of information content in MFA could help crack this shortcoming. Nonetheless, the question remained abstract and it was not clear what a sound solution could look like.

The task to develop a context-specific understanding of phenomena such as "information", "data quality" and "uncertainty", which typically are devoid of general and clear definitions, was accompanied by the incentive to formally approximate quantitative metrics for objects that lack statistical information. The multifacetedness of the research required including concepts from a wide range of research fields. Great colleagues at TU Wien provided important reflections and impulses at critical points of this study and helped to open my eyes for neighboring scientific disciplines. As we will see later, ideas from information sciences all the way to theoretical ecology find their place within the proposed concept and a useful metric for system structure is formally related to the information content measure.

The result of this effort is put together in this thesis. It builds upon three journal articles:

> Schwab, O., O. Zoboli, and H. Rechberger. 2016. A Data Characterization Framework for Material Flow Analysis. *Journal of Industrial Ecology*.
>
> Schwab, O., D. Laner, and H. Rechberger. 2016. Quantitative evaluation of data quality in regional Material Flow Analysis. *Journal of Industrial Ecology*.

Schwab, O. and H. Rechberger. Information Content, Complexity and Uncertainty in Material Flow Analysis. *Journal of Industrial Ecology*. Under revision.

I primarily contributed to the three articles. This includes conceptualization and formalization of the methodology, analysis of the case studies, contextualization and discussion of the research, and preparation of the articles. Helmut Rechberger provided impulses to all three articles, Ottavia Zoboli contributed to the case study of the first article, and David Laner contributed to the mathematical procedures and the case study presented in the second article.

An MFA-specific approach to information is presented in Chapter 2. It is organized according to the three articles listed above and provided in appendix 8, including identical text blocks, tables, figures and appendences. Applications of the proposed methodology to regional MFAs are presented in chapter 3. Beforehand, general perspectives on MFA, uncertainty and information are provided in Chapter 1.

Being aware that the topic addressed in this thesis yields the potential for virtually endless debate and refinement, I hope the reader finds the solution proposed here interesting to follow and also useful to apply.

Acknowledgements

I would like to express my great appreciation to my supervisor Helmut Rechberger for many stimulating discussions and ideas which helped shaping this thesis, and to Reinout Heijungs and Gang Liu for their feedback and evaluation. I am thankful to Mag. Dr. Robert Holnsteiner of the Austrian Federal Ministry of Science, Research and Economy, for providing funding for important parts of this work. I am grateful to the reviewers and editors at *Journal of Industrial Ecology* for the very valuable comments to, and for the assistance in the publishing process of, the three papers which form the basis for this cumulative thesis. Special thanks for inspiring discussions and invaluable impulses go to my colleagues Nađa Džubur, Ottavia Zoboli, David Laner and Oliver Cencic, and to Inge Hengl for her graphical support. Heartfelt thanks to the whole team of the research group at TU Wien for contributing to such a great working environment, and to the colleagues in Vienna and in other parts of the world for the great times we were privileged to share. You made my time as doctoral candidate special and unforgettable. With my warmest and deepest gratitude, I thank all which I gratefully call my loved ones, for enriching my life and for their ongoing support.

TABLE OF CONTENTS

1 Introduction

1.1 Material Flow Analysis and regional metabolism

Ecology has a long tradition in describing material flows in ecosystems, such as nutrient flows in food webs or carbon flows spanning biological and physical environments ("carbon cycle"), as consequences of natural processes. With increasing mobilization, transformation and use of materials by societies, the influence of anthropogenic activities as drivers of material flow systems increases (Klee and Graedel 2004; Baccini and Brunner 2012). Growing awareness of the limits to growth, of resource depletion and of environmental degradation are advancing a field of research concerned with understanding the flows of materials within systems that span natural and anthropogenic environments. This field of research is referred to, for example, as *metabolism of cities* (Wolman 1965), *industrial metabolism* (Ayres 1994), *metabolism of the anthroposphere* (Baccini and Brunner 1991) or *society's metabolism* (Fischer-Kowalski 1998). Essentially, it refers to the idea of approaching industrial systems as if they were ecological systems (Frosch and Gallopoulos 1989), which is today a central concern of the field of *Industrial Ecology* (Graedel 1996).

Within Industrial Ecology, Material Flow Analysis (MFA) is a widely applied analytical tool for modeling, understanding and optimizing material flow systems by comprehensive investigation of material flows into, within and out of a given system. MFAs include sets of processes, which are defined as transformations, relocations or storages of materials. These processes are, according to the specifications made by an MFA modeler, connected via flows. A process that stores a material includes a so-called material stock, which may increase or decrease over time, depending on the balance of all input and output flows of the respective process. MFAs provide useful information on metabolic systems that span natural, technological and economic environments. Procedures for preparation of MFAs and for representation of MFA results have been largely harmonized (Baccini and Bader 1996; Brunner and Rechberger 2004; ASI 2005) and material flow studies today are of increasing size and level of detail (see, among many others, Reck et al. (2010), Liu and Müller (2013), Nakajima et al. (2013), Habib et al. (2014)). MFAs of various scopes and materials have proven useful not only in scientific discourse but also in decision making in policy and industrial contexts (see, for example, Vadenbo et al. (2014), Trinkel et al. (2015), Zoboli et al. (2016), Hofko et al. (2016)).

A typical goal of an MFA is to understand the metabolism of a specific material within a region, such as a nation, by means of detailed analysis of the supply, consumption and disposal of this material within a defined time interval, usually one calendar year (see, for example, Egle et al. (2014), Bonnin et al. (2013), Figure 1).

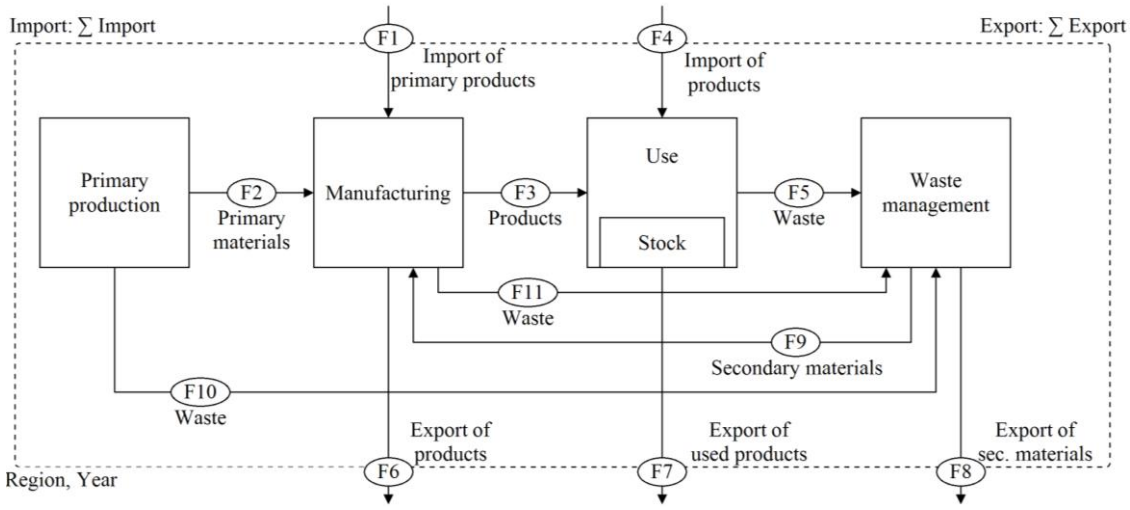


Figure 1: A generic national material flow system including import and export flows, and flows between primary production, manufacturing, use and waste management. Arrows represent flows, boxes represent processes and the broken line represents the system boundary. In the system illustrated, the process "use" includes a stock. More detailed MFAs may contain diversified sets of flows and processes.

A national material flow system usually covers not only imports and exports of an economy, but includes also more specific information on material flows within the economy. National MFAs are investigated in more detail later in this thesis (for an example, see Figure 4 on page 15). The structure and level of detail of national material flow systems depend mainly on the goal of the MFA, on author choices and on the available data basis (Klinglmair et al. (2016)). Frequently, however, the information available for description of material flow systems is considerably limited and, as a consequence, data uncertainty considerations become increasingly relevant.

1.2 Uncertainty

Uncertainty in science may relate to context definition, model structure, model inputs, parameter values, and others (Walker et al. 2003). As a general concept, it is proposed to distinguish *epistemic* uncertainty and *aleatory* (also stochastic, or natural) uncertainty (Morgan et al. 1992). Epistemic uncertainty is understood as uncertainty due to limited or imperfect knowledge, which could be reduced by further investigation. Aleatory uncertainty is understood as uncertainty due to natural

variations or randomness, and it cannot be reduced. While epistemic uncertainty relates to knowledge shortcomings, aleatory uncertainty refers to the impossibility of reducing certain entities (or objects, phenomena) to simple empirical quantities such as one precise value. In that sense, variability is an intrinsic property of any entity that has more than one realization. It is thus also an intrinsic part of a complete piece of information. Not knowing about the extent of variability, in return, is a knowledge shortcoming and thus epistemic uncertainty.

An uncertainty type of central interest in this work is data uncertainty. Ideally, data uncertainty can be understood as a problem of variability and can, if sufficiently large datasets are given, be quantified by statistical methods. These possibilities may, however, be limited as given data may not always be sufficient for proper application of statistical methods. Moreover, data uncertainty is not always a unidimensional phenomenon but may also, in addition to variability, include elements of epistemic uncertainty such as disagreement, linguistic imprecision, systematic error or subjective judgement, and others (Morgan et al. 1992). Recognizing the subjective element in uncertainty, it can be regarded as relating to the degree of confidence an agent has in certain outcomes or probabilities (Refsgaard et al. 2007). In the course of this thesis, uncertainty in MFA is identified as a phenomenon depending on data quality, where data quality is understood as a multidimensional and partially subjective phenomenon, defined as the degree of belief an agent has in given data being true in a particular context. Understanding uncertainty as to refer to imperfect or missing information, uncertainty is formalized as the counterpart of information later in this thesis (paragraph 2.3).

1.3 Information

Scientific activity has always been undertaken with the aim of revealing or creating information, be it by observing and describing phenomena such as nature in its immediate surroundings, the relation of objects in space or the behavior of individuals, among many others. Plato probably presented the first comprehensive work on the phenomenon "information" in his "Theory of Forms", where he correlated concepts such as *observation*, *knowledge*, *memory* and *idea* and established the terminological ground of *information* as knowing the *form* or structure of an object (Adriaans 2013). Since then, an ever increasing body of information has been formed by the sciences. Today's possibilities to observe phenomena, to manipulate data and to multiply information are as manifold as never before. Concurrently, the number of concepts on information has increased: Popular concepts define information as a process (Capurro and Hjørland 2003), as a quantity (Fisher 1925;

Shannon 1948; von Neumann 1955; Kolmogorov 1968) or as a state (Hintikka 1973). A generic philosophical approach to information was formulated by Floridi (2013), who says that "information is data with meaning". This approach is revisited later in this thesis (paragraph 2.1).

One of the most widely applied formal concepts of information is information theory, as coined by Shannon (1948). Here, *information* can be understood as opposed to *uncertainty*, in that additional information reduces uncertainty, and *vice versa*. In mathematical resemblance to the thermodynamic concept of entropy (Clausius 1867; Boltzmann 1872), Shannon formalized information entropy as the expected value of the information contained in a message. Information entropy is a quantity referring to unpredictability or uncertainty in an event or a set of events. Information theory is the central basis of a concept from theoretical ecology on which the approach proposed in paragraph 2.3 expands.

While information theory provides useful approaches to technical aspects of information (such as communication and transmission), it is limited regarding the semantic content of information, that is, its meaning. Shannon and Weaver (1963) ask "How precisely do the transmitted symbols convey the desired meaning?" and referred to this "semantic problem of information" as being relatively more intricate than the sheer engineering aspects. Indeed, the meaning of any piece of information depends on the understanding of the agent processing this piece of information, and thus information always has a subjective element (Arndt 2004). Moreover, statistical analysis of data may contain subjective elements (Berger and Berry 1988). De Finetti (1974) argues in his "Theory of Probability" that there are only subjective probabilities, where probability is "the degree of belief in the occurrence of an event attributed by a given person at a given instant and with a given set of information". In addition to subjectivity, problems of unclear semantics have been observed as a key limitation for making good use of data (Madnick and Zhu 2006). It appears that "information" depends, similar to probability in the conception of de Finetti, also on perspective, semantics and context. These perceptions of information are revisited later in this thesis (paragraph 2.2).

1.4 Material Flow Analysis, information and uncertainty

Although studies of material flow systems can provide information, they also depend on information in their production process, and a lack of useful information can be a limiting factor to the level of detail provided in an analysis. More than that, the results are typically inherently limited in terms of accuracy and, thus, in their reliability in subsequent decision-making processes (Graedel et al. 2004;

Chen and Graedel 2012). Despite the importance of data quality for the validity of results, it is not always clear how data shortcomings are reflected in MFA results (Chen and Graedel 2012). Clearly, if MFA is seen as a way of compiling data to create information about material stocks and flows and to aggregate this information to create knowledge about material flow systems, the quality of its fundamental components, data, is substantial. Recognizing the shortcomings of MFA data in combination with the variety of sources and the various ways collected data are applied in the analysis process furthers appreciation of the fact that the databases of studies are not always comprehensible for agents other than the producer. Essentially, there is no collective understanding about what data or, more generally, information in MFA is and how it can be characterized.

MFAs are often based on cross-disciplinary, highly heterogenic data. These data may have different formats and qualities and come from heterogeneous sources, such as official trade statistics, scientific literature, consumer behavior studies and expert estimates. In many cases, MFA data are not based on empirically well-founded datasets, but on isolated values which are not always provided in consistent formats. In some cases, extensive statistical data, such as lab data on substance concentrations, might be available, but analysts usually have to cope with isolated values. Consequently, statistical methods of data uncertainty evaluation are often inadequate in MFA practice (Hedbrant and Sörme 2001). Additionally, relevant data may be confidential, lost, highly aggregated, or outdated, or real-world phenomena may be too complex to be directly measured and information must be derived in other, indirect ways. Furthermore, the background of data is not always transparent because of missing meta-information. Data may be inaccurate due to measurement and collection errors, be biased by the interests of data producers, or be unrepresentative and incomplete (see Table 1). Data quality shortcomings have, besides model uncertainty (uncertainties due to simplifications and assumptions in model design), been identified as major sources of uncertainties in environmental modeling (Björklund 2002), also in MFA (Danius 2001). Since established scientific methods are often limited when it comes to uncertainties, alternatives to traditional problem-solving strategies are required (Funtowicz and Ravetz 1993). This motivates also the application of non-traditional strategies for dealing with uncertainty in environmental analysis and assessment methods (Heijungs and Huijbregts 2004).

The evaluation of data quality and the treatment of data uncertainty have been addressed in different areas that model environmental systems (Refsgaard et al. 2007). In Industrial Ecology, established statistical procedures such as stochastic modeling and scenario modeling are often applied for the

treatment of uncertainties, for example in input-output models (Lenzen et al. 2010), in Life Cycle Assessment (Lloyd and Ries 2007) and in MFA (Gottschalk et al. 2010). Often, these approaches require more information than is actually available as data are typically given in the form of individual, isolated values and not in the form of statistically exploitable datasets. This holds especially for MFAs such as the case studies presented later in this thesis, where data uncertainty relates to knowledge shortcomings ("epistemic uncertainty", Laner et al. (2014)). Consequently, even though there are methods for treatment of known data uncertainties in MFA (see, for example, Kopec et al. (2015) and Cencic (2016a)), means for actual characterization and representation of data uncertainty in the absence of statistical evidence are limited.

Table 1: Perspectives on MFA data and requirements of MFA data quality (collected from MFA modelers at TU Wien in an internal workshop in December 2014)

Requirement	Description	Complement
The data exist.	The data have been collected personally or by another agent.	non-existence
The data are available.	Data collected by another agent are provided or communicated.	non-availability
The meaning of data is clearly defined.	The entity and the respective data about it are precisely defined, unambiguous in its meaning and linguistically precise ("semantic precision").	ambiguity
Data is provided at a sufficient level of detail.	Available data are detailed enough, i.e. the resolution of the data is high enough for the desired context.	high aggregation
Data for a model are complete.	There are enough (consistent) data to exactly determine or over-determine and thus to balance the system.	incompleteness
Data on an (extensive) entity are complete.	Extensive (system-size dependent) entities are the result of a summing process and are complete when all required components are considered in this process.	fragmented data
Data on an (intensive) entity are representative.	The data at hand are representative for the entity studied, i.e. the number of samples is adequate to specify the entity studied. This also relates to "completeness".	unrepre-sentativeness
The data fit the system boundaries and the context.	The applied data are within the temporal and spatial system boundary, and describe the entity of interest.	inadequacy
The data producer is known and reliable.	The source of the data is known and considered reliable.	unreliability of producer
It is known how the data were created.	The formation process of the data is transparent and can be reproduced.	unknown origination
Data can be cross-checked.	Data can be compared to semantically similar data from an independent reference and thus be verified.	non-verifiability
Meta-information is sufficient for data quality evaluation.	The provided meta-information about the data is sufficient for data quality evaluation.	non-transparency
Information on uncertainty is provided.	Quantitative information on data uncertainty is available.	no information on uncertainty

As alternatives to author judgements or expert estimates of uncertainties (as, for example, performed in Graedel et al. (2004), Huang et al. (2007) and Ott and Rechberger (2012)), more systematic and

transparent approaches have been proposed. In a concept of Hedbrant and Sörme (2001), MFA data are assigned to five uncertainty levels according to their origin, and this classification is then translated to uncertainty ranges. Expanding on that idea and integrating elements of the LCA-specific data quality concept of Weidema and Wesnæs (1996), Laner et al. (2015b) propose a concept in which data uncertainty ranges are formalized as functions of five data quality indicators.

Material flow modelers often choose to represent data uncertainty by means of uncertainty ranges, also because these can be treated in established frameworks. If no information on statistical variability is provided, however, the idea behind uncertainty ranges is to express the degree of belief an agent has in given data to be true, although it may be difficult to specify uncertainty ranges in the absence of empirical evidence. As a consequence, besides the choice of distribution geometries, the specification of uncertainty ranges will probably be arbitrary: Why is an uncertainty range of ± 20% for quantity *X* assumed and not a range of ± 30%? Is ± 100% a natural upper limit, or is that often rather chosen because of mathematical convenience and the physical constraint that the lower bound of a quantity is zero? The question remains whether it is useful to quantify the unquantifiable, that is, to provide quantitative uncertainty ranges when these are actually unknown, also because this conveys the unjustified impression of empirical evidence. With the incentive to avoid the use of uncertainty ranges but to still allow for both relative and absolute comparisons, in this thesis, uncertainty is regarded as a system property of MFAs which involve imperfect information. As formalized later in this thesis, the potential uncertainty of a system increases with its size (its number of flows) and it decreases when more and better data are incorporated. Approaching uncertainty as the counterpart of information (see paragraph 1.2), the information content of an MFA increases when the system uncertainty decreases, and *vice versa* (see paragraph 2.3).

1.5 Objectives

As elaborated by Klinglmair et al. (2016), the reliability of MFA results and their system structure often depend on probably subjective choices, limited information, and the structure of available data. Comparisons of material flow systems regarding the reliability of their results and their system structures are, however, restricted to mainly qualitative considerations. To provide a quantitative basis for future evaluations and comparisons of MFA systems is the aim of this study. In this thesis, a data characterization framework for MFA is proposed. A formal procedure for the estimation of data quality based on data characteristics is presented, and a system-theoretical approach to quantitative

evaluation of system uncertainty, information content and system structure in MFA is proposed. Although this thesis focuses mainly on the role of quantitative information (data) in MFA, qualitative information necessary for composing qualitative MFA models is also an element of the system-theoretical approach. The thesis culminates with a set of procedures and metrics useful for evaluating and comparing the databases of MFAs, their data qualities, and their information contents and system structures. The usefulness of these measures is illustrated in four case studies (phosphorus, palladium, plastics and aluminum). Parts of these case studies are used to exemplify the procedures and measures elaborated in the following methodological chapter.

2 Methodology

With the aim of providing a quantitative basis for evaluation and comparison of MFA systems, a methodology consisting of four steps is proposed in this chapter (Figure 2).

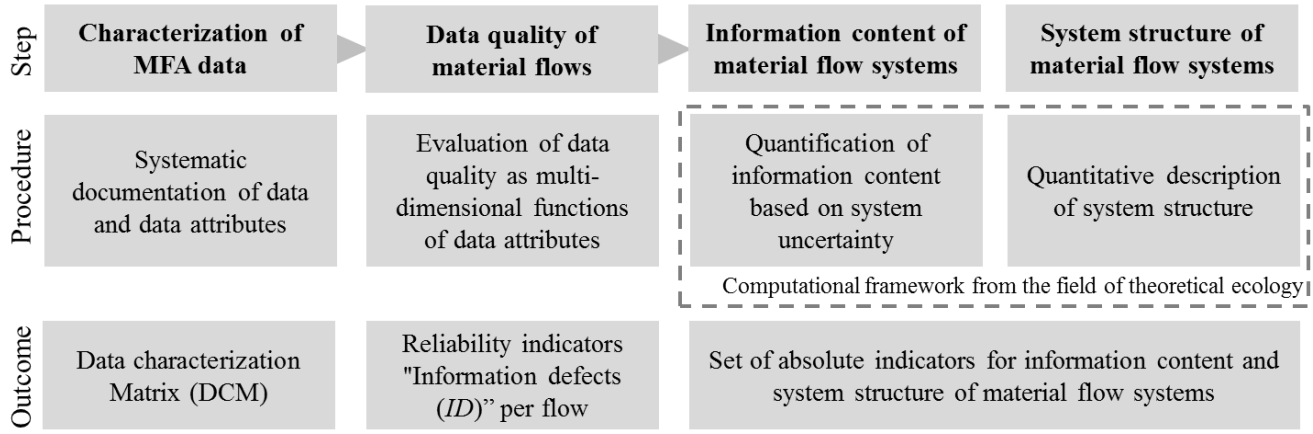


Figure 2: Outline of the four linked methodological steps presented in this chapter. The first two steps (data characterization and data quality evaluation) are prerequisites of the third step (quantification of information content). The third and the forth step (quantification of information content and quantitative description of system structure) are computed in a network-analytical framework adapted from the field of theoretical ecology. Outcomes of the proposed methodology are a comprehensive documentation of MFA meta-data and a set of absolute measures for the information content and the system structure of material flow systems.

A framework for characterization of MFA data is presented in paragraph 2.1, a formal procedure for evaluation of data quality is presented in paragraph 2.2 and a procedure for the quantitative evaluation of information content and system structure of material flow systems is presented in paragraph 2.3. A focus of the presented methodology is on flows, that is, on their contribution to the information content of and their role in the structure of MFA systems.

2.1 MFA data characterization

In this paragraph, a framework for consistent description and characterization of *a priori* MFA data (before application in a model) is presented. This is the basis for analysis of an MFA study's database structure, and for data quality evaluation. The proposed procedure is illustrated by application to a regional MFA of phosphorus. The benefits and shortcomings for MFA practice are discussed. The core of this framework is a data characterization matrix (DCM) which facilitates the systematic documentation and characterization of MFA data. Before the DCM is introduced, central terms are defined.

2.1.1 Terminology

This terminology is to provide a conception of data and information in MFA as a basis for precise communication within and beyond the research community, and to contribute to a common understanding of quantitative information in MFA. The terminology is the foundation of the data characterization framework.

2.1.1.1 MFA system elements

MFA system elements are the components of material flow systems, i.e., "flows", "processes", "stocks", and "materials" (Brunner and Rechberger 2004). "Flows" are specified as mass per time, processes as dimensionless transfer coefficients, and "stocks" as mass. "Material" is an umbrella term for goods and substances. Each system element is assigned a specific number as an identifier, i.e., a flow or process number. Cross-boundary flows (flows that leave or enter the system) are called imports and exports, and flows within the system (between processes) are called internal flows. One or more related processes and associated flows can be referred to as "sectors", such as "industry and trade sector" or "consumption sector". Designating sectors can improve the comprehensibility and ease of communication about material flow systems. It also enables comparing systems that differ in their overall composition of processes but consist of similar sectors.

2.1.1.2 Entity, data element and attribute

An entity is a real-world phenomenon or real-world object, and its realizations are represented as data. If data in MFA are considered quantitative information, data are representations of entities as numeric values (see Floridi (2013)). That is, an entity can be represented by a data element (isolated value, interval or dataset). The number of a study's data elements can be larger than the number of entities as more than one reference could be available for quantification of an entity (for example

three independent references on a phosphorus concentration of an agricultural good, that is, three data elements on one entity). The total of all data elements per entity is referred to as information element. MFA data attributes are data-associated annotations concerning statistical properties, meaning, origination and application of the data. Attributes can be designated as the "characteristics of data" (Wang et al. 1995) and specify a data element, the relation to the entity it represents, its origination and formation process and its relation to the application context.

2.1.1.3 Information level

Four levels of information in MFA can be distinguished (Figure 3). The first information level is "data element", as described above, and a data element plus meaning forms "information" (Floridi 2013). "Information background" represents the origination and forming process of the piece of information. Placed in context, it forms "MFA information".

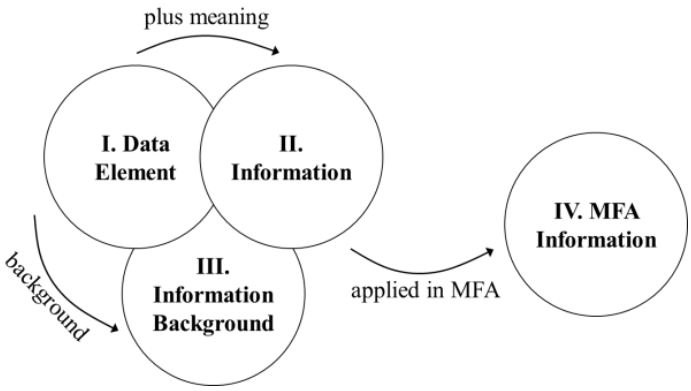


Figure 3: MFA information is information in MFA context: A data element plus meaning forms information, this information has a background and in the context of an MFA study it forms MFA information.

For example, the entity "aluminum content of a beverage can" is to be specified for an MFA study. The datum is, say, "95". This forms information, with its meaning "aluminum content of a standard beverage can in central Europe in 2010, in %". The "information background" is, for example, that it has been measured by an academic research group by x-ray analysis, but the specific observation method and the number of samples are unknown. It forms "MFA information" when applied in an MFA study as a material specification for a designated flow or stock. MFA information can be described by sets of attributes that are arranged according to the four distinct information levels, as proposed in the below introduced data characterization matrix.

2.1.1.4 Data semantics

Semantics refers to the "intrinsic meaning" of a piece of information, and data that is meaningful and truthful can become information (Floridi 2013). For example, the data at hand may describe the "phosphorus content of national annual crop production in 1990". This specification of the datas' meaning lacks semantic precision, as the notion of "crop" is ambiguous. It is not known whether it refers to food crops, to cereals, or also to energy crops and industrial crops. Data semantics can also change over time (Madnick and Zhu 2006), such as when the variety of cultivated crops changes. Unclear data semantics can lead to data misinterpretation and, consequently, to drawbacks in data quality.

2.1.1.5 System relation and system adequacy

System relation refers to the sphere that determines the data (such as market processes, technological state-of-the-art, biosphere) and the variability of a datum over time, space and other potential relations. Other relations can be, among others, technology (for example productivity rates can differ between production plants) or reference units (such as data that refer to fiscal years instead of calendar years). MFA data should be adequate for the studied system with respect to time, space and potential further relation. For example, data from a neighboring country might be temporally adequate but spatially inadequate, or might be inadequate as they describe a different technical process (further relation).

2.1.1.6 Autonomy and application of data

Data that can be directly introduced in a model for the description of system elements are referred to as to be autonomous in their application. Often, there are no ready-to-apply autonomous data available for the description of MFA system elements. These need to be instead quantified by the combination of several non-autonomous data elements. Data elements can be applied in an MFA study as one of the typical application types (flow, flux, stock, transfer coefficient and material). Other data elements such as areas and numbers are summarized as precursors. For example, readily applicable data of mineral phosphorus fertilizer use from consumption statistics, given in mass of phosphorus per year, is autonomous for the purpose of a national phosphorus MFA. In contrast, the flow of phosphorus in animal manure (flow, t/yr) is non-autonomous if it needs to be calculated from the number of animals (precursor, dimensionless), excretion per animal type (flux, kg/animal·yr) and the phosphorus concentration of animal excrement (material, %). The more non-autonomous data

elements there are to be combined for the description of a system element, the higher is the number of potential data quality impairments.

2.1.1.7 Origination of MFA data

Data for MFA can be acquired either from direct observations, such as measurements, monitoring or counting ("empiricism"), or can be abstracted from given information. In contrast to empiricism, the latter is in this context referred to as "derived" and is divided into three categories: "mainly from data" (such as reporting data that is aggregated by statistical offices), "mainly from assumptions" (such as data from models with many assumptions because of a scarce database), and "from speculation" (such as guesses).

2.1.1.8 Variety and disparity

The attributes "variety" and "disparity" describe the complexity of a population. Variety refers to the number of potential real-world objects an entity refers to, disparity to the spread of these real-world objects' realizations. For instance, "copper content of smartphones" can refer to a vast number of different smartphones (high variety) and the copper content of these smartphones can span a wide concentration range (high disparity). In contrast, both the variety and disparity of the "aluminum content of aluminum cans" are comparably small, as the number of different types of aluminum cans is limited and the range of the aluminum content is rather narrow (between 95 and 99%). A more precise specification of a data element's meaning (for example, to a particular type of smartphone) can reduce variety and disparity. Data quality can decrease because of improperly understood data semantics and limited context knowledge (Madnick and Zhu 2006) and that information always has a subjective element (Arndt 2004). This is considered in the data characterization framework, which at the same time is designed for a high degree of transparency and replicability. Key to the framework is the characterization of MFA data by specification of data attributes in a data characterization matrix (DCM).

2.1.2 MFA data characterization matrix (DCM)

The database of a material flow system is documented, structured, and analyzed in the DCM. The DCM has been developed in an iterative process by the analysis of several regional MFAs (Schwab and Rechberger 2014). In the matrix, 49 data attributes are assigned to each data element of a study. The DCM is structured according to the four information levels (Figure 3) and related attributes are grouped in attribute groups (Table 2).

Table 2: Structure of the data characterization matrix by information levels and attribute groups

Info. level	Attribute group	Description (no. of attributes)	Attributes
Data element	Statistical characteristics	Documentation of statistical information on a data element (10).	Data element form, location parameter, value (numeric), n, min, max, distribution (form), distribution (paramet.), dispersion (measure), dispersion (numeric)
Information	Semantics	Specification of the meaning of a data element (2).	Description of meaning, semantic precision
	Scale	Specification of the format of an entity (8).	Entity category, entity class, unit, sphere, property type, mathematical form, min (potential), max (potential)
	Complexity	Description of the complexity of an entity (2).	Variety, disparity
Information background	Availability	Distinction if wanted information does exist and is accessible or not (3).	Existence, accessibility, access restriction
	Communication	Documentation of how a piece of information is communicated (3).	Communication type, access type, frequency
	Producer	Documentation of the agent that produced the piece of information, for example an authority (3).	Producer category, producer type, reference
	Origination	Documentation of the data collection method, for example counting or industrial monitoring (3).	Origination category, origination type, origination type quality
MFA information	Application in MFA	Description of how a piece of information is applied in the MFA study (4).	Application type, autonomy, layer, type of good
	System relation	Description of the relation between a piece of information and the studied system (6).	Primary determination, temporal variability, trend, spatial variability, further relation, variability by further relation
	System adequacy	Description of a piece of information's adequacy to (resp. divergence from) the studied system (5).	Temporal divergence, spatial divergence, further divergence, adaptation (type), adaptation (quality)

A more detailed description of each data attribute is provided in appendix 2. For application of the DCM to a given MFA database, each of these data attributes is specified individually. For specification of the attributes, a code has been developed. By this code, attributes are assigned to

particular measurement scales (absolute, nominal, binary, ordinal) and ranges of possible data attribute specifications are provided. This facilitates the consistent completion of the matrix, also when applied by different researchers to different regional MFAs, and enables automated analysis of a DCM once completed. The DCM code and examples of completed matrices are provided in appendences 3 and 1.

In the following, the data characterization framework is illustrated in a database analysis of a national MFA. This database analysis consists of three steps, which are (a) creation of data inventory, (b) evaluation of data elements, and (c) analysis of data attributes. In (a), all system elements and the respective data elements are listed in the DCM. In (b), the attributes are specified with the help of the DCM code, and in (c), the DCM is analyzed attribute wise.

2.1.3 Application of the data characterization framework

The data characterization framework is applied to the 2009 phosphorus system of Austria (Zoboli et al. (2015), see Figure 4), which is based on the work of Egle et al. (2014). A comparatively sound database for quantification of material flows and stocks of this phosphorus MFA is available. Data uncertainties were assessed by an approach by Laner et al. (2015b) and range from 10% to 90%. Nine out of ten flows have less than 40% uncertainty, and two-thirds of the flows have less than 30%. These relatively low uncertainties (compared with other regional MFAs) underline the database's robustness.

Key information on the system and the applied database is provided in Table 3. The phosphorus MFA of Zoboli and colleagues is a flow-based model in which the number of applied transfer coefficients is kept to a minimum. Respectively, the scope of the here presented case study is limited to the evaluation of these phosphorus flows in the main system and does not include processes and subsystems.

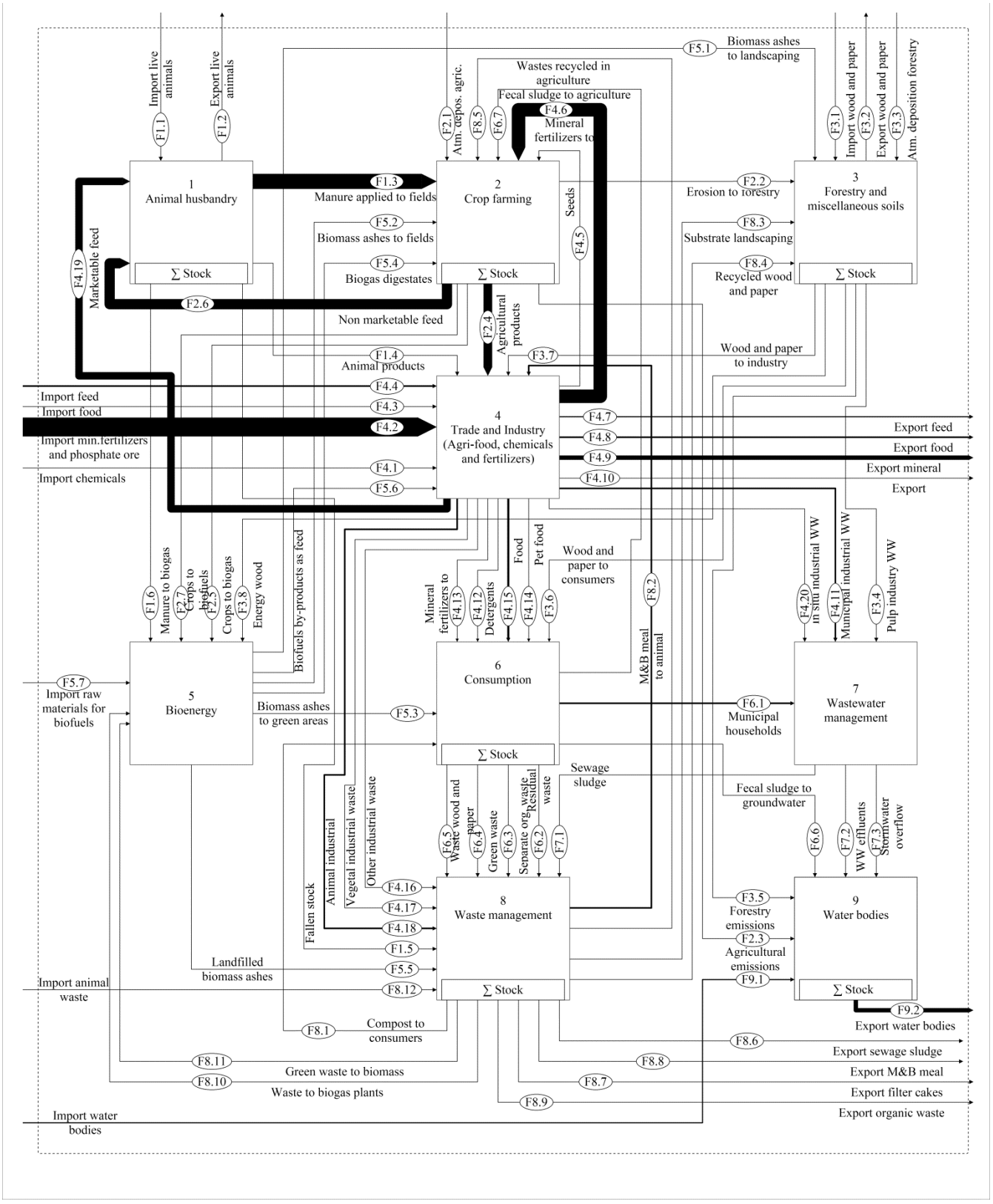


Figure 4: Flowchart of the Austrian phosphorus MFA according to Zoboli et al. (2015).

2.1.3.1 Data inventory

The total of 308 data elements and all assigned data attributes are inventoried in the DCM (see supporting information of Schwab et al. (2016a)). As listed in Table 3, these 308 data elements are used for the description of 172 entities and are aggregated for the description of 72 flows. Twenty percent of these flows are quantified directly by autonomous data and 80% by the combination of data on two or more entities.

Table 3: Key information on the structure and the data basis of the 2009 Austrian phosphorus MFA

database characteristic	quantity
number of flows in main system	72
number of processes	9
number of subsystems	8
number of stocks	7
total number of collected data elements	308
total number of entities	172
average entities per flow	2.4
average data elements per entity	1.8
share of flows that can be described directly by autonomous data (%)	20
isolated values (%)	75

2.1.3.2 Evaluation of data elements and analysis of data attributes

The elements of the data inventory are evaluated by specification of data attributes according to the code for data characterization (Schwab et al. 2016a). Exemplarily, selected attributes are analyzed in the following: data producer (Figure 5a), data origination (Figure 5b), utilization type (Figure 6a), entity class (Figure 6b), type of good (Figure 7) and primary determination (Figure 8). Please note that the quantities given here are not material quantities but "information quantities". The number of samples *n* in figures 3-6 are relates to the number of collected data elements (n =308) or the number of entities (n=172).

More than half of the data were collected from authorities, about 40% from scientific sources (Figure 5a). Generally speaking, data on material flows stems from authorities and data on material qualities (composition) from science. Approximately 40% of the data elements are from empirical collections (such as measurement or counting) and 55% are derived (either from data, assumptions or speculations). Most prominent are the reported data from third parties that are aggregated by

authorities (Figure 5b) such as official trade statistics. These contribute to the generally more robust database for cross-boundary flows (for example for imports of goods) in contrast to the often weaker database within the system (for example in the consumption sector).

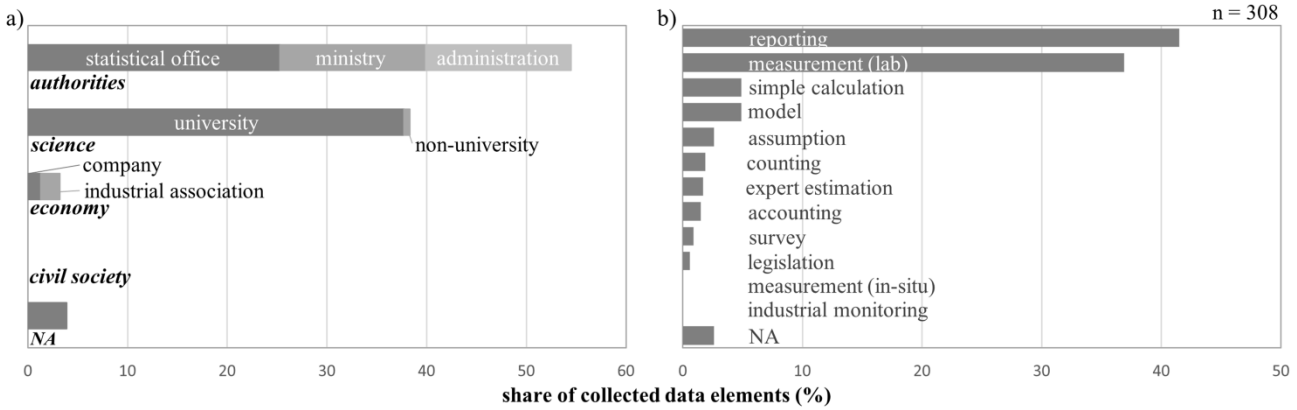


Figure 5: Producer category and producer type (a) and origination type (b) of data in the phosphorus case study.

The two most prominent references, namely, reporting data from statistical offices and empirical data from scientific measurements, are complemented by data from additional sources. Expert estimations are important especially in the consumption and waste management sectors, assumptions in the bioenergy sector, and scientific models in the waste management and crop farming sectors. For animal husbandry, simple calculations based on data from authorities and science complement directly applicable data. More than 40% of the data are communicated in reports, 35% in online databases, and 10% in scientific journals or books (cf. attribute no. a305 (access type) in the supporting information of Schwab et al. (2016a)). The number of data elements per entity (on average 1.8, see Table 3) is less than or equal to four in 95% of the cases, and 75% are isolated values.

Most of the collected data describe material flows (Figure 6a) and come in the format “mass/time” (Figure 6b). Approximately one-sixth of the collected data are precursors mainly on numbers and areas and need to be combined with other data before introduction to a model.

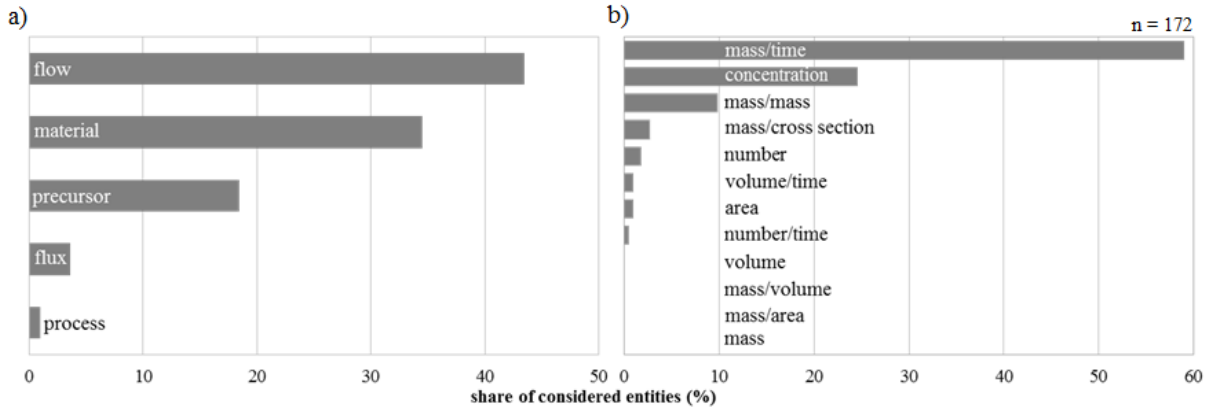


Figure 6: Utilization (a) and entity class (b) of data elements. (a) describes the utilization type of data used in the study and (b) the format of the collected data. "Concentration" is in mass-%; mass/mass refers to other entities such as productivity rates.

Forty percent of the collected data describe waste, 25% consumer goods, and 20% industrial goods (Figure 7). The label "none" refers to other entities, such as conversion factors or areas. Although the waste management sector has less flows than other sectors, such as industry (see Figure 4), most of the collected data are on waste. This indicates that in this case, less directly applicable, autonomous data for the description of the waste management sector is available and in consequence the overall data search effort is greater.

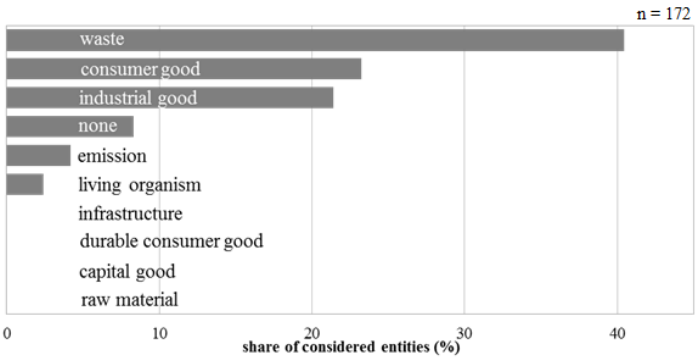


Figure 7: Collected data relating to different types of goods. "None"= no goods but other entities such as areas or conversion factors.

The attribute "primary determination" (Figure 8) refers to the spheres that primarily determine the data values. For example, the phosphorus concentration of common wheat is primarily determined by the biosphere and phosphorus removal rate of a sewage plant by the applied technology. In the analyzed study, 40% of the data elements are primarily determined by market activities (such as domestic production of agricultural goods), 10% by technology (for example, phosphorus removal rate from wastewater), 8% within the sociosphere (for example, consumer behavior), 30% in the

biosphere (such as phosphorus content of crops), and 6% in the geosphere (such as discharge of rivers per time unit). Data that is primarily determined by political decisions is applied mainly in the waste management and bioenergy sectors (for example amount of phosphorus in fecal sludge applied on agricultural fields). Examples of applied scientific rationales are molecular masses of phosphorus and phosphorus compounds.

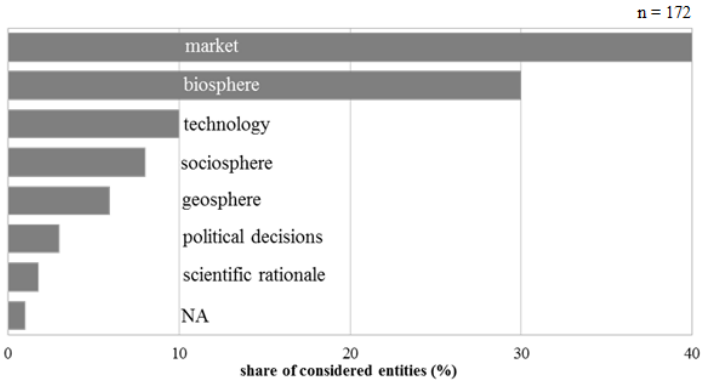


Figure 8: Primary determination (mechanisms that primarily determine the value of data and their change over time) of data within the anthroposphere and the natural environment. "NA" = not available.

Although the primary determination of data is not always unequivocal, it can indicate the main factors that shape the data of an MFA and thus the material system itself. Clearly, the quantity of applied data elements does not necessarily correlate with the physical quantity of the material flows of a study. However, this can contribute to the identification of a system's main driving phenomena. The driving factors of material systems in terms of physical quantities have been investigated by Klee and Graedel (2004). Transferring this idea from physical quantities to information quantities, the DCM can be used to reveal the driving mechanisms of a material system from its database. The database structure of the case study indicates that, regarding its information quantities, the Austrian phosphorus MFA appears to be strongly perturbed by anthropogenic activities but to be not entirely dominated as there is still a prominent influence of the natural environment (i.e., biosphere and geosphere, see Figure 8) on the material system.

In the phosphorus case study, statistical offices (aggregated reporting data on material quantities) and scientific literature (measurement data on material qualities) are the central data sources. Data from industry and also from civil society (for example from interest groups) are not applied in the phosphorus study. The database is found to be better for cross-boundary flows than for flows within the system. This is especially because of detailed official foreign trade statistics. In contrast, institutionalized statistics such as the latter are limited within the system, for example in the use

sector. The results indicate that there is a general tendency of the databases to become weaker from upstream to downstream sectors, i.e., from primary production and manufacturing to waste management. Especially in use and in waste management, data producers are required to be more active in providing disaggregated and transparent data in consistent formats.

2.1.4 Discussion of the data characterization framework

The data terminology can be the basis of coherent data communication across different studies and research groups. The DCM facilitates the systematic documentation of MFA data and designated attributes. Attribute-wise data evaluation draws a compact picture of this database as illustrated in Figure 5 - Figure 8. The figures promote a simple and condensed representation of MFA databases, for instance, in reports or publications, as an alternative to communicating extensive data tables. Systematic tagging of MFA data with data attributes can further the understanding of the information basis of a study, enables comparing different MFA studies to one another and can give indications regarding the quality of the data. Nevertheless, it has to be considered that the database is subject to the scope and level of detail of a study, which is determined by the focus of the research. In the analyzed material flow system, processes such as crop farming are rather treated as "black boxes", while wastewater and waste management processes were ranked higher in the specific interest of the research group and were thus studied in more detail. From the experience of this study, it can be said that the quality of available information decreases when moving downstream the material flows (see also Mao et al. (2008), Graedel et al. (2004)). Previous database analyses indicated the tendency of decreasing data quality when moving from the main system into specific subsystems with higher level of detail (Schwab and Rechberger 2014b). Both tendencies are also due to a decreasing share of empirical data from authorities and science, an increasing share of speculations and expert estimations, and the decreasing autonomy of the data. The net working time for a database analysis of a study with the extent of the above described MFA is approximately 60-80 hours. In further research, it is recommended that the DCM is applied simultaneously to the data collection process rather than posterior to an MFA.

In the case study, meta-information on the meaning and the formation process of data was sometimes found to be limited, although it is imperative for data producers and data publishers to provide this information. From the experience of this database analysis, it can be said that meta-information may be lost or become imprecise in the scientific publication and citing process. Over time, data can

appear “just to be there”, without precise knowledge about its initial meaning and collection method, which might lead to poor data quality estimations and poor application of the data. Moreover, data can be misinterpreted not only due to ambiguous data semantics but also because of diverging reference units. This was found to be the case in the phosphorus flow system, as most of the data refer to calendar years (1 January – 31 December), but some do refer to fiscal years (often from June to June).

Attribute specifications are, in part, subject to authors’ judgments. Therefore, these specifications can be argued over by third persons with diverging perspectives. Author judgments on MFA data have also been an element of previous studies (Graedel et al. 2004; Hedbrant and Sörme 2001). The novelty of the approach presented here is that it is not the data as such that are judged by the authors, but individual data characteristics, (data attributes). Inevitably, subjectivity is intrinsic to these judgments. Nevertheless, subjectivity is also intrinsic to information *per se*. Ignoring the subjective part of information can restrain a comprehensive understanding of it (Arndt 2004; Berger and Berry 1988). For these reasons, subjectivity is also an element of the here-presented framework, even if it is controlled by a standardized and transparent procedure, which, however, can facilitate a discourse about MFA databases and collective learning of material flow systems.

2.2 Data quality evaluation

In this chapter, a phenomenological model for quantitative evaluation of data quality is presented. “Phenomenological model” is understood in a sense similar to McMullin (1968), where a phenomenon which cannot be directly observed (data quality) is approximated based on observable phenomena (data attributes). Data quality is regarded as the degree of belief in data to be true in a given context, and this degree of belief is influenced by data characteristics that are believed to be relevant. Accordingly, data quality is expressed as a multi-dimensional function of data attributes selected from the set of data attributes proposed in paragraph 2.1. As a result, each flow of a material flow system can be described by a value indicating the degree of belief in data to describe a flow of interest truthfully. This value is calculated based on evaluation of all data elements applied for quantification of a particular flow (see Figure 9). It is thus specific to any material flow of interest within a temporally and spatially defined system.

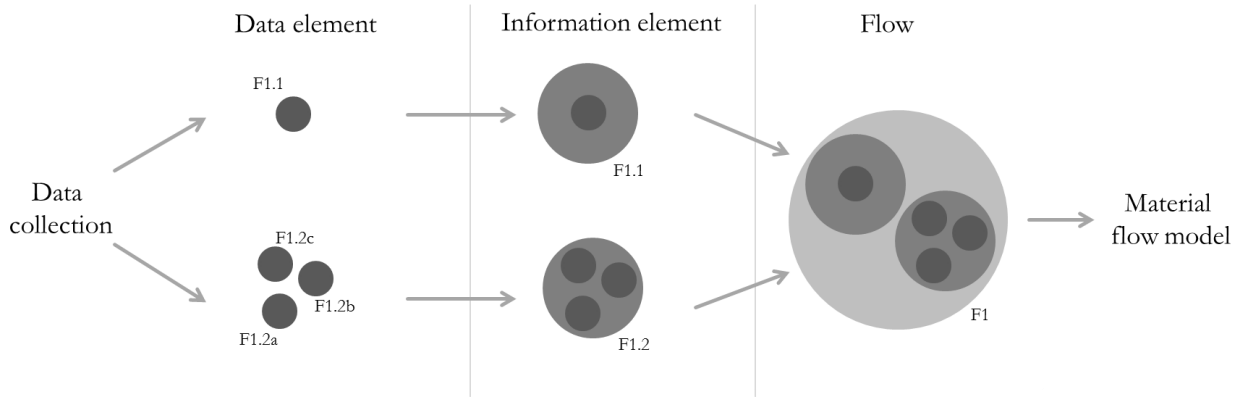


Figure 9: A typical quantification of a material flow "F1". F1 consists of two information elements, which itself consist of one (F1.1) or more (F1.2) data elements.

A flow may be described based on two information elements. An information element itself can be specified by one or more data elements, where data elements are representations of real-world objects or real-world phenomena ("entities") as numerical values (isolated values, intervals or datasets). With regard to Figure 9, F1.1 may be information on the number of cars imported into an economy, provided by official trade statistics. F1.2 may be information on their Pd concentration in %, provided as an expert estimation in personal communication (F1.2a) or as measurement results communicated in scientific literature (F1.2b and F1.2c). The data quality of flow F1 depends on the data quality of all data elements used for its quantification. The hierarchy from data elements to information elements to flows is a core premise of the data quality evaluation procedure proposed in this paragraph. For illustration, the procedure is explained in detail for one flow of the Austrian palladium (Pd) MFA (Laner et al. 2015a) in the following. Consequently, the procedure is applied to all flows of the Pd MFA and its information basis is evaluated. The benefits and shortcomings of the presented procedure are discussed.

2.2.1 Approach and conceptualization

Uncertainty in regional MFA is rather an epistemic than an aleatory phenomenon, i.e not a consequence of natural variability but of imperfect knowledge. Knowledge shortcomings are here expressed as "defects of information" (Dubois and Prade 2010). Information defects are belief indicators that reflect the deviation of given information from a desired state of perfect knowledge. They are expressed on an ordinal scale from 0 (no information defect) to 1 (maximum information

defect). The four information defects “semantic”, “representativeness”, “provenance” and “context” (ID_S, ID_R, ID_P, ID_C) appear to be relevant for regional MFAs (Figure 10).

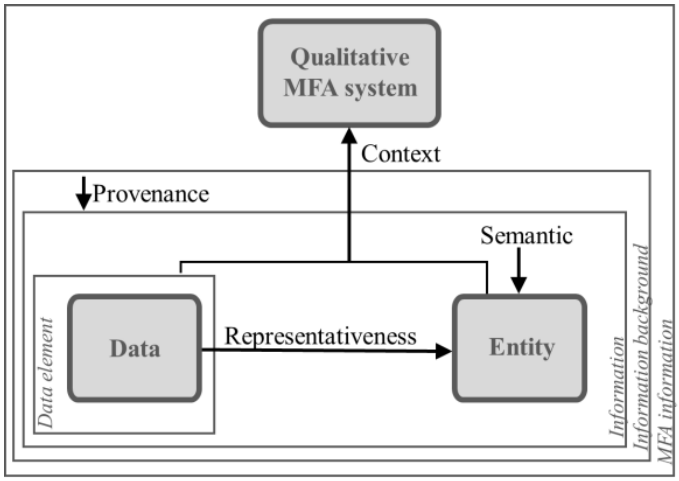


Figure 10: Concept of MFA information defects and their position in the data characterization framework presented in paragraph 2.1. "Data" are numerical values, "entity" is a real-world object or phenomenon described by an information element, "qualitative MFA system" is a system to be quantified by introduction of quantitative information.

ID_S refers to the semantic precision or resp. imprecision of the meaning of data (Does the specification “smart phones” also refer to mobile phones from before the technological leap, which are still “out there”?). ID_R indicates how well a given data element represents the entity of interest (Is the complex entity “Pd concentration of mobile phones” quantified based on one or more measurements or independent references?). ID_P considers the origination and collection method of a data element (How reliable are the information producer and the data collection method?). ID_C designates how well a given data element fits the context of a study (Is the data element timely and does it refer to the geographical area studied?). These information defects are to some degree similar to data quality indicators found to be useful in previous studies (Laner et al. 2015b) in terms of, for example, the correlation in the dimensions “time”, “space” and “further”, which are here part of the context information defect ID_C. Other defects, such as the semantic information defect ID_S or the representativeness information defect ID_R, are new concepts of the approach presented here.

The approach is illustrated in detail for flow ten (F10) of the palladium case study (see paragraph “Case study on Pd flows in Austria 2011” Laner et al. (2015a)). A fully characterized information inventory of the Pd flow study is provided in appendix 1. Flow F10 refers to the Pd flow in flat screens sold in Austria in 2011. Laner and colleagues quantified flow F10 based on two information elements. First, this is the per capita flow of flat screens (information element F10.1) and second, the Pd content in flat screens (information element F10.2). F10.1 was calculated from data on the 2010

German market and related assumptions. For F10.2, information from a scientific report providing German data of the year 2010 was used.

The information defects illustrated in Figure 10 are exemplarily qualified for information element F10.2 (Pd content in flat screens). The information element F10.2 has a semantic information defect (ID_S) because "flat screens" is not a clear specification as there are different types of flat screens. F10.2 refers to a complex entity as different types of flat screens (attribute "variety") may also differ in their Pd content (attribute "disparity"). This complex entity is quantified based on one reference. A complex entity in combination with a small number of references or samples induces a representativeness information defect ID_R. Because no information is provided on the data collection method, there may also be a provenance information defect (ID_P). The data do not fit the actual system context (Austria 2011) as they are for Germany in 2010. Consequently, there is a context information defect ID_C.

This vague qualitative description of information defects enables first estimates on the overall quality of the data. A mathematical formalization of the qualitatively introduced concept of information defects is proposed in the following.

2.2.2 Formalization

The information defect per flow ID_F is quantified in three steps. First, the quality of each data element is described by a set of four defects ID_i (ID_S, ID_R, ID_P and ID_C, see Figure 10). Second, each information element is described by one total information defect (ID_{tot}), which is an aggregation of the ID_i of the respective data element. Third, the data quality of each flow is described as ID_F, which is a combination of the ID_{tot} of the respective information elements (according to the order illustrated in Figure 10). Prior to data quality quantification, the database of an MFA study has to be inventoried and characterized according to the data characterization framework (Schwab et al. 2016). The procedure of data quality quantification is described in the following. Exemplarily, the information defect of the flow F10 of the Pd MFA introduced above is quantified.

2.2.2.1 Data attributes

Quality-relevant data attributes are listed in Table 4 and exemplarily specified for the information elements F10.1 and F10.2 in the rightmost columns. Data attributes in text format (for example the producer type, which may be, among others, "national statistics" or "industrial association") have been translated to mathematically computable formats according to a translation scheme provided in

appendix 4. Consequently, all attributes in Table 4 are specified either on an ordinal scale from 0 to 1 where 0 means "good" and 1 means "bad", on an absolute scale (0, 1, 2…), or on a binary scale (0 means yes, 1 means no).

Table 4: Data attributes relevant for data quality evaluation selected from the data characterization matrix introduced in paragraph 2.1.2 and their attribute numbers as identifiers. The designators are used in the proposed formal procedure. Attributes are exemplarily specified for information elements F10.1 and F10.2 in the rightmost columns according to the code provided in appendix 3

Data attribute	Attribute no.	Designator	Scale	F10.1	F10.2
Number of samples	*a104*	n	Absolute	1	1
Semantic precision	*a202*	a	Ordinal	0.3	0.2
Variety	*a211*	b	Ordinal	0.8	0.5
Disparity	*a212*	c	Ordinal	0.4	0.7
Producer type	*a308*	d	Ordinal	1	0.3
Origination type	*a311*	e	Ordinal	0.4	0.4
Origination quality	*a312*	f	Ordinal	0.7	0.5
Temporal variability	*a406*	g	Ordinal	0.2	0.2
Spatial variability	*a408*	h	Ordinal	1	0.1
Variability by further relation	*a410*	i	Ordinal	0	0
Temporal divergence	*a411*	j	Ordinal	1	0.1
Spatial divergence	*a412*	k	Ordinal	0.1	0.2
Further divergence	*a413*	l	Ordinal	0.2	0.0
Adaptation type	*a414*	m	Binary	0	1
Adaptation quality	*a415*	o	Ordinal	0.3	0

In the first step of the evaluation procedure, the four information defects ID_i are quantified based on the 15 attributes listed in Table 4.

2.2.2.2 Information defects of data elements (ID_i)

The four information defects ID_i (ID_S, ID_R, ID_P, ID_C) are described as functions of data attributes (Table 4). The information defect functions have been developed in a two-step heuristic procedure. First, the basic function type was chosen. Second, the relationship between data attributes, as qualified in chapter "approach and conceptualization", was formalized as combination of data attributes by use of the chosen function type. The designators *a, b, ..., o* of the attributes (Table 4) are used in the functions presented in the following.

ID_S is regarded as a linear function of the attribute "semantic precision" (*a*, see Table 4, where a=0 represents data with unambiguous and clear meaning and a=1 represents data with ambiguous or vague meaning), which means that the information defect is high when the meaning of data is vague (Eq. 1).

$$ID_S = a \quad \text{Eq. 1}$$

ID_S of information element F10.2 equals the data attribute "semantic precision", so that $ID_{S,F10.2} = 0.2$.

The representativeness information defect ID_R is formalized as an exponential function of the attributes "variety" (*b*) "disparity" (*c*) and "number of samples" (*n*). ID_R increases with increasing variety and disparity (that is, with increasing complexity of the described entity). ID_R and the information gain per additional sample decrease with increasing numbers of samples (Eq. 2). This relates to the equation of the standard error of the mean (SEM), where the error (expressed as standard deviation) decreases with increasing sample size (see Clark-Carter (2014)).

$$ID_R = (\sqrt{b}\sqrt{c})^{n/(n+1)}\left(\frac{\sqrt{b}\sqrt{c}}{\sqrt{n}}\right) \quad \text{Eq. 2}$$

The information element F10.2 refers to a complex entity with high variety and disparity and a small number of samples (see Table 4), so that $ID_{R,F10.2}$=0.46.

The provenance information defect is formalized as a function of the information producer (attribute "producer type" (*d*), first term in Eq. 3) and the way the data were collected (attributes "origination type" (*e*) and "origination quality" (*f*), second term in Eq. 3). The exponents determine the slope and the curvature of the function, i.e. their non-linearity. An exponent >1 results in a convex curved function. This means that ID_P is high only if both the information producer and the data generation method are specified with high attribute values. This way, data of a "bad" data producer is not *per se* evaluated as "bad" (as would be the case in a concave function, that is, with an exponent <1) as long as a "good" data generation method was applied. Here the exponents of the first and the second term are defined identically, that is, the information producer and the collection method have the same relative weight (Eq. 3).

$$ID_P = \left(d^{1.5} + \left(\frac{e+f}{2}\right)^{1.5}\right)/2 \quad \text{Eq. 3}$$

F10.2 was collected from a reputable scientific report which provides expert estimations on substance concentrations. The provenance information defect is, based on the attributes listed in Table 4, $ID_{P,F10.2}$=0.23.

The contextual information defect is formalized as a product of two constitutive parts. First, this is the degree to which data fits the system studied. Second, this is the quality non-adequate data were adapted to the system, for example by scaling. In Eq. 4, *y* denotes the data adequateness (ordinal (0-1), see Eq. 5), *m* designates if data was adapted (binary, yes/no, resp. 0/1) and *o* refers to the adaptation quality (ordinal, 0-1).

$$ID_C = y - (1 - m)(1 - o)y \qquad \text{Eq. 4}$$

This means that non-adequate data (expressed as *y*, first term) causes a high information defect, which decreases if this non-adequate data was well adapted to the context (second term).
The variable *y* in Eqs. 4a and 4b denotes the degree to which data does not fit the context in three dimensions: time, space and further (such as technology). It is a function of the divergence of the data from the system context ("divergence in three dimensions", *j, k, l*) and the variability of the data ("variability in three dimensions", *g, h, i*).

$$y = (\sqrt{g}\sqrt{j} + \sqrt{h}\sqrt{k} + \sqrt{i}\sqrt{l})/3 \qquad \text{Eq. 5}$$

F10.2 does temporally and spatially diverge from the system boundary ("divergence"), but is little variable over time and space ("variability") as the composition of flat screens in Austria and Germany can be regarded as quite similar in two subsequent years. The data has not been adapted to the system boundary. Considering the data attributes listed in Table 4, it is $ID_{C,F10.2}$=0.09. The information defect functions are visualized as surface plots provided in appendix 5.

2.2.2.3 Information defects of information elements (ID_{tot})

The four information defects are aggregated to a total information defect per information element (ID_{tot}) as Euclidian distance (the shortest connection of any point to the origin in an n-dimensional space) in a four-dimensional space. This is normalized to the measurement scale (0-1) by the number of information defects ID_i=4 (Eq. 6).

$$ID_{tot} = \sqrt{\frac{(ID_S^2 + ID_R^2 + ID_P^2 + ID_C^2)}{4}} \qquad \text{Eq. 6}$$

In Eq. 6, ID_i are weighted by themselves. That means that an information element with a high defect in one dimension cannot be of overall good quality, even if the other three defects are low. When applied to F10.2 ($ID_{S,F10.2}$=0.20, $ID_{R,F10.2}$=0.46, $ID_{P,F10.2}$=0.23, $ID_{C,F10.2}$=0.09), this results in $ID_{tot,F10.2}$=0.28. The procedure until here can be repeated for information element F10.1 so that $ID_{tot,F10.1}$=0.42. The defect of information element F10.1 (number of flat screens sold) is higher than the defect of information element F10.2 (Pd content in flat screens). The information defect of Flow F10 can be expressed as one flow-specific value by combination of the two total information defects $ID_{tot,F10.1}$ and $ID_{tot,F10.2}$.

2.2.2.4 Information defects of flows (ID_F)

ID_F is formalized as the square root of the sum of squares of all ID_{tot}, analogous to the combination of uncertainties in the Gaussian rule of error propagation (see the exponent in the denominator in Eq. 7, where z designates the number of information elements). This term can potentially increase indefinitely for increasing z and must consequentially be normalized to the measurement scale (0-1). Realistically, z is virtually never higher than four (a substance flow is typically quantified by multiplication of two information elements, quantity of goods times concentration, and in fewer cases by multiplication with additional information such as, for example, on volumes or areas). The term could be normalized by the square root of the numbers of information elements ($\sqrt{z}$). This straight forward normalization is not sensitive to the number of information elements z and it averages the information defect of multiple information elements. However, it appears to be more suitable to consider that the more imprecise information there is to be combined, the less credible is the result. Having that in mind, ID_F can be also normalized by applying a logistic function such as the one proposed in Eq. 7. This function accumulates the information defects of multiple ID_{tot} per flow. A graphical comparison between normalization by $\sqrt{z}$ and a logistic function is provided in appendix 6.

$$ID_F = \frac{1.5}{1 + 2e^{-3\sqrt{\sum_{i=1}^{z} ID_{tot,i}{}^2}}} - 0.5 \qquad \text{Eq. 7}$$

Applied to flow F10 with $ID_{tot,F10.1}$ = 0.42 and $ID_{tot,F10.2}$=0.28, this is ID_{F10} = 0.54.

2.2.2.5 Information elements specified by more than one data element

The evaluation procedure has been illustrated for the situation of one data element per information element. As illustrated in Figure 9, information elements may also be quantified based on more than

one data element. For example, the information element F1.2 of the Pd study (Pd content of cars imported to Austria) consists of three data elements (Table 5 and appendix 1). That is, reference A states that the Pd content is A%, reference B says B% and reference C says C%. Apparently, agents frequently introduce the mean of available data elements in their model when they cannot discriminate between the reliability of the three references.

In case of more than one data element per entity, ID_P and ID_C are calculated on the level of data elements and ID_S and ID_R are calculated on the level of information elements (i.e. per entity). That is, because each data element may have a different provenance (different ID_P) and may be of different adequateness to the context (different ID_C), but is used to represent the same entity (same ID_R) with the same meaning (same ID_S). This becomes clear when reconsidering the concept presented in Figure 10.

Table 5: Information element F1.2 of the Pd study consists of three data elements. The lowest ID_P and ID_C are selected for further processing in Eq. 6

Information element	Data element	ID_S	ID_R	ID_P	ID_C
F1.2		0.20	0.23		
	F1.2a			0.25	0.05
	F1.2b			0.28	0.22
	F1.2c			0.23	0.18
$ID_{i,F1.2}$		0.20	0.23	0.23	0.05

Eq. 6 requires a set of four information defects ID_i per information element. To formulate such a set of four ID_i for the situation presented in Table 5, a straight forward approach is chosen. Experience shows that typically, an agent only introduces additional data elements per information element when expecting information gain (for example by taking the mean of two independent references). This is considered here, and the lowest ID_P and ID_C are selected from the set of provenance and context defects (see Table 5, where F1.2 is quantified based on n=3 data elements). Consequently, the information defect $ID_{tot,F1.2}$ decreases (because of min ID_P and ID_C and also because n>1 in ID_R), which reflects the information gain.

The information defect approach results in a new quantity for evaluation of regional MFAs. This new quantity indicates the reliability of model input data and enables distinguishing material flows by their data quality. The evaluation procedure is applied to all flows of the Pd MFA in the following.

2.2.3 Application of the information defect approach

The presented procedure for quantification of data quality is applied to the palladium (Pd) flow system illustrated in Figure 11. For a more detailed description of the Pd MFA and a quantitative diagram, please refer to the article of Laner et al. (2015a).

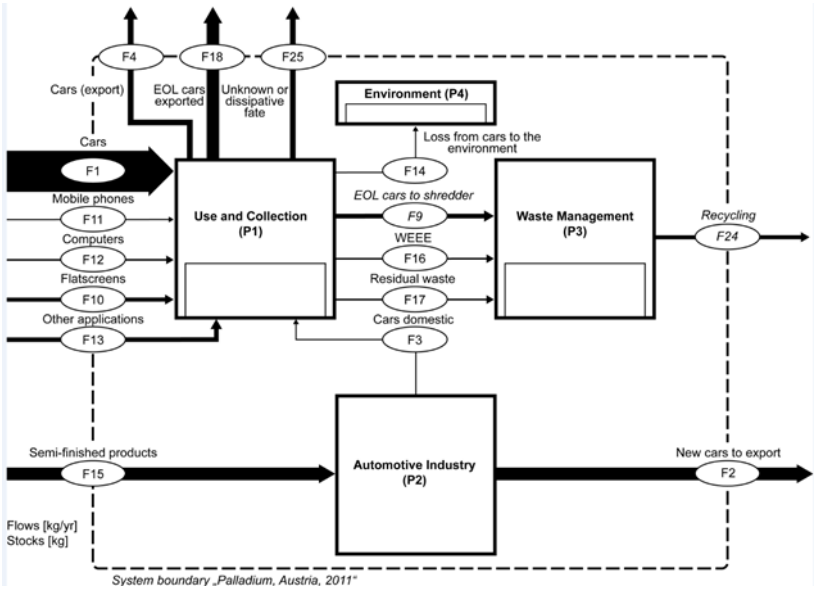


Figure 11: Structure of the 2011 Austrian Pd MFA by Laner et al. (2015a). The system consists of 25 flows (16 in the main system and 9 in subsystems "use and collection" and "waste management").

Information defects ID_i per data element, ID_{tot} per information element and ID_F per flow are computed according to Eqs. 1-6. The results are illustrated in Figure 12. A detailed table of all information defects is provided in appendix 1.

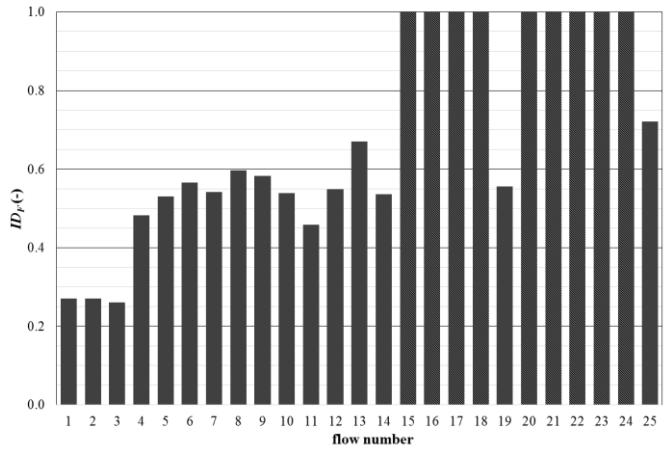


Figure 12: Data quality of the flows in the Pd MFA expressed as information defects ID_F. Low defects indicate good data quality, high defects indicate poor data quality. Flows without input data are here assigned $ID_F=1$.

With regard to the concept of information defects, low bars indicate good data quality (flow F1 – F3) and high bars indicate poor data quality (flows F4-F14, F19, F25). Information defects higher than 0.5 signify data of considerably poor quality. For some flows, no input data were available. Clearly, non-existent information cannot be defective, but complete ignorance can be regarded as a maximum information defect. Thus, ID_F=1 is assigned for unknown flows (flows F15-F18, F20-F24). The bars in Figure 12 denote the *a priori* knowledge about flows, that is, the knowledge before application of a material flow model. By balancing of flows in a model (in the Pd study, the STAN algorithm (www.stan2web.net) was applied), initially unknown flows are calculated and the *a posteriori* state of information differs from the *a priori* information state. Until here, the information defects enable assessing the state of information about a material flow system and to underpin qualitative observations about available information by quantitative means (subsequent applications of information defects in material flow modeling are outlined in the "concluding remarks" section). For example, data quality is often found to decrease over the lifecycle of materials and to be better for sectors of economic interest such as trade and manufacturing in contrast to the consumption and waste management sectors (see, among others, Mao et al. (2008), Graedel et al. (2004)). The results of the Pd case study indicate that data quality is considerably higher at the system input side, while data quality of flows to the environment (such as dissipative fate, flow 25) is poor. For many flows in the waste management sector (for example flows 20, 21), no data are available. The information defects do now provide an opportunity to quantitatively express data quality and to illustrate weaknesses and tendencies of the database in a systematic and reproducible way.

2.2.4 Discussion of the information defect formalization procedure

Data attributes can be mathematically combined in many different ways for specification of information defects. The formalizations of ID_i, ID_{tot} and ID_F proposed here deliver mathematically sound and reasonable results for quantitative data quality evaluation. They have been selected from a number of possible formalizations based on comprehensive qualitative and quantitative tests, where individual steps of the quantification procedure have been varied and compared regarding their absolute output and their relative ranking based on Monte Carlo Simulation, surface plots and correlation analysis (Schwab and Rechberger 2015). The mathematically simplest alternative approach is to formalize the defect of information elements as an average (denoted as $ID_{tot,average}$ in the following) of all ordinal attributes (Table 4). In Figure 13, this $ID_{tot,average}$ is plotted against ID_{tot} of the information elements of the Pd case study.

The averaged information defect appears to equalize the results and to deviate from ID_{tot}, especially for increasing information defects. Although $ID_{tot,average}$ is mathematically feasible, it is of little meaning with regard to the information defects. That is because some data attributes are obviously related to others, which is not considered by $ID_{tot,average}$. For example, the attribute "temporal divergence" interacts with the attribute "temporal variability" when it comes to data quality evaluation as it is obvious that outdated data (temporal divergence) is only defectuous if the data varies over time (temporal variability). The example of $ID_{tot,average}$ indicates that the adequateness of very simple mathematical formalizations to express information defects may be limited.

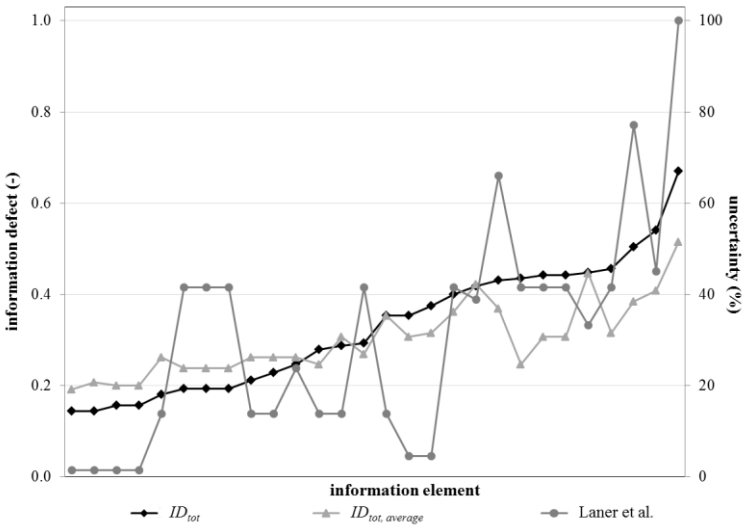


Figure 13: Comparison of ID_{tot} to an alternative total information defect $ID_{tot,average}$ and data uncertainty estimations of Laner and colleagues. Information defects (dimensionless) are plotted on the primary, uncertainties (%) on the secondary y-axis. The values are sorted according to increasing ID_{tot}. The connecting lines between the points are introduced to simplify visual comparison of the plotted options.

To assess and discuss the results of the data quality approach, the information defects of the Pd case study are also compared to data uncertainties as calculated in Laner et al. (2015a) (see Figure 13). Laner and colleagues used an adapted version of Hedbrant and Sörme (2001) for calculation of data uncertainties in the Pd study. While this is based on categorization of data into five quality categories according to their origination, the information defect approach distinguishes data quality by a number of data characteristics and considers interconnections between data attributes. Figure 13 indicates that uncertainty calculations and ID_{tot} can differ, but show a similar trend (Spearman rank correlation coefficient between uncertainty ranges and ID_{tot} in the comparison presented is ρ=0.7, between ID_{tot} and $ID_{tot,average}$ it is ρ=0.8). The range covered by ID_{tot} seems less wide than the range covered by the uncertainty estimates (Figure 13).

A difference between the introduced method and other approaches to data quality, such as the one in Laner et al. (2015b), is, that data quality is not formalized based on static indicators and categories. Data quality is here formalized in a model-type setup, where different data characteristics are linked and may enhance or reduce the resulting information defect, depending also on the magnitude of related attributes. The data attributes contribute to the information defects ID_i, ID_{tot} and ID_F to a variable extent, depending on the model formalization. The weight of data attributes in ID_F in the formalization proposed above has been analyzed. This was done by investigating the relative impact of variations in inputs (individual data attributes, Table 4) on the observed variation of the output (ID_F, Eq. 7) in a sensitivity analysis (Monte Carlo-based multiple linear regression). The relative weight of data attributes is displayed in Figure 14.

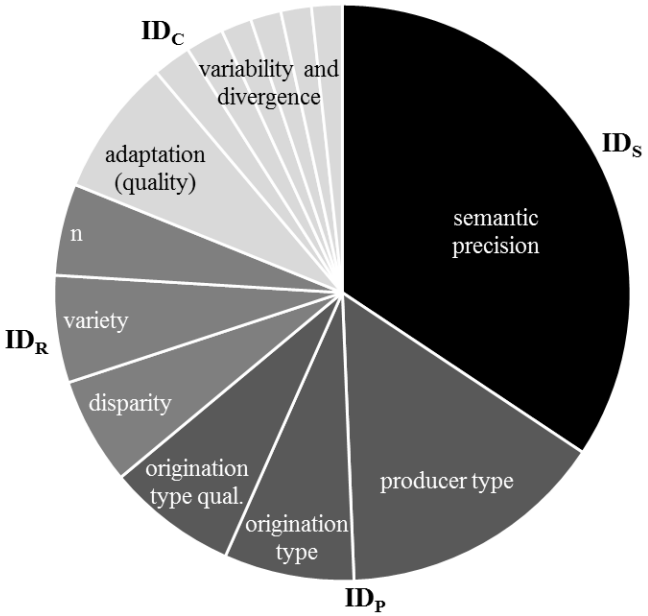


Figure 14: Relative weight of individual data attributes and ID_i in information defect ID_F.

According to the formalization presented here, the precise knowledge of the meaning of data (ID_S) and the provenance of data (ID_P) contribute most to ID_F. In contrast, all input attributes considered would have the same weight in the alternative formalization $ID_{tot,average}$ mentioned earlier. Certainly, however, the weight of data attributes in the information defects can be varied, for example, by the introduction of weighting factors in Eqs. 5 and 6, or by modification of Eqs. 1-4.

The application of the data quality evaluation procedure may require more time than other approaches (see the overview of existing methods in Laner et al. (2015b)). In return, it enables better

understanding of the factors determining the information quality of material flows. For convenient und time-saving application, the model presented here is implemented in a spreadsheet tool attached to the data characterization matrix (paragraph 2.1). In that spreadsheet tool, data quality can be calculated automatically once a database has been characterized. Evaluation of a study with the extent of the Pd case presented here requires 30-40 work hours. More detailed full-scale national material flow systems (the Pd study was not full-scale, it had a focus on EOL of consumer product flows) may require more time for their characterization. It may be difficult to retrieve all information necessary for sound data characterizations of existing studies that used other data documentation schemes or that provide no complete and consistent data documentation. Thus it is beneficial to apply the data characterization framework while preparing a study, that is, in parallel to data collection.

Objective (statistical) possibilities for data quality evaluation are usually limited in MFA. Data quality evaluation is inevitably subjective, also with the proposed procedure. Different authors, also with similar backgrounds and perspectives, may still propose differing assessments of data quality. Nevertheless, the proposed procedure reduces the influence of author opinions and subjectivity by systemizing the evaluation procedure and by moving choices from generic data classifications to an evaluation of individual data attributes. Despite the systematic approach, agents with different backgrounds and with different degrees of knowledge may characterize data and their attributes differently and may thus produce differing results in data quality evaluation, especially when the approach is applied not in parallel but subsequent to a study. The method presented should be seen as a transparent "best guess" procedure for facilitating reproducible and transparent expert estimations on the abstract phenomena of "data quality".

The output of the method presented is a ranking of flows on an ordinal scale according to their data quality. An alternative concept of Laner et al. (2015b) aims at providing uncertainty ranges for MFA data. Often, the initial idea behind uncertainty ranges is to express the reliability of the underlying data. Naturally, it may be difficult to express epistemic uncertainties ("lack of knowledge") as absolute measures and the definition of uncertainty ranges may be highly speculative. That is why the approach presented here is designed to evaluate the degree of credibility of *a priori* data without simulating an absolute quantification of uncertainty. Information defects can in subsequent work be applied as dimensionless factors in characterization functions of material flow models and as indicators for the reliability of data (for example in Laner et al. (2015a)) or as factors in data reconciliation algorithms (for example Kopec et al. (2015)). The information defects can be applied

as indicators for epistemic uncertainty in data uncertainty frameworks (such as, among others, Dubois and Guyonnet (2011) and Clavreul et al. (2013)).

Uncertainty ranges are practicable and often desired by material flow analysts. The dimensionless information defects can also be translated to uncertainty ranges by application of scaling functions and by multiplication with a coefficient of variation or an uncertainty factor. More than that, it is possible to test whether the integration of empirically derived probabilities into the information defect concept is adequate for particular MFA applications. As statistical characteristics such as dispersion measures or probability distributions are, if available, also part of the data characterization matrix (see attribute group "statistical characteristics" in the data characterization matrix presented in paragraph 2.1, this could also be exploited for characterization of observed variability. As the absence of information on variability can also be referred to as epistemic uncertainty, it is also part of the approach presented here (see the representativeness information defect (ID_R) in Eq. 2, which increases with decreasing number of samples n).

In combination with the data characterization framework presented in paragraph 2.1, the evaluation procedure proposed enables the documentation, characterization, evaluation and communication of the information basis of regional MFAs. The information defects indicate the reliability of data and help to find weak points in the data structure. They enable identifying the reason for data weaknesses (Is the source unreliable? Is the number of samples not high enough? Is the meaning of the data unclear?) and aid in adopting adequate measures for filling data gaps. When not interpreting information defects as factors for uncertainty evaluation but leaving them as dimensionless measures for a "state of knowledge", they can be applied for measuring the information content of MFAs and for comparing MFAs of different substances, regions or years to one another, as proposed in the following paragraph.

2.3 Information content and system structure

From a graph-theoretical perspective, MFAs can be regarded as networks. These networks can be represented as directed graphs, in which flows connect source processes to target processes. Typically, network structures in Industrial Ecology can be distinguished by their topologies, and trivial structures can be distinguished from non-trivial, or complex, structures. In MFA, "complexity" is frequently understood as a structural feature of systems helpful for distinguishing material flow systems, or for expressing that systems are perceived to be similar (see, for example, the comparison of two national phosphorus systems in Klinglmair et al. (2016)). In this paragraph, the structural

"complexity" of material flow systems is, as opposed to their structural "triviality", addressed in the same computational framework as the information content and the uncertainty of material flow systems.

It has been argued that, in Industrial Ecology, uncertainty and complexity issues are increasingly relevant (Kay 2002) and that information is a notable phenomenon, also because "uncertainties paralyze us" and because it is not clear how much information is needed for design of systems (Bettencourt and Brelsford 2015). This holds also for MFA: When recalling that studies of material flow systems both reveal new information and depend on existing information when prepared (Chen and Graedel 2012), the dual role of information in MFA becomes apparent. For making informed MFA-based decisions, agents are not only to know about the actual MFA results on a material level, but also about their reliability, that is, about the „uncertainty" and, respectively, the „information content" of a given material flow system. For making good use of available data, agents also are to know about the "complexity" of a material flow system, as increasingly complex systems require, in comparison to systems of more trivial structures, increasing amounts of information in order to be solved. In this paragraph, the phenomena uncertainty and complexity are addressed as properties of regional material flow systems. The information content of material flow systems is derived from their uncertainty.

2.3.1 Uncertainty in Material Flow Analysis

As elaborated in paragraph 1.4, empirical evidence for evaluation of uncertainties in MFA is typically limited and as a consequence, established statistical measures are limited in their applicability. In the same regard, uncertainty ranges are of limited meaning for representation of data uncertainties, although they are frequently applied in MFA. As an alternative to the approaches of Hedbrant and Sörme (2001) and of Laner et al. (2015b), the concept of information defects (*ID*) has been introduced in paragraph 2.2. Information defects are indicators for the belief of degree an author has in data to be true in a specific context. Low information defects ($ID \rightarrow 0$) relate to data of good quality, high information defects ($ID \rightarrow 1$) relate to data of poor quality and $ID=1$ relates to complete ignorance. The *IDs* are central variables for consideration of data quality in a quantitative approach to uncertainty and information content in MFA as presented in this paragraph. The incentive is to avoid the use of uncertainty ranges, but to still allow for absolute evaluations and comparisons. As formalized later in this paragraph, the potential uncertainty of a system increases with its size (its

number of flows) and it decreases when more and better data are incorporated. Uncertainty is approached as the counterpart of information (cf. paragraph 1.2) and the information content of an MFA system increases when the system uncertainty decreases, and *vice vers*a.

2.3.2 Complexity in Material Flow Analysis

Complexity concepts are of increasing interest for system analysis in Industrial Ecology, as put together in two special issues of the *Journal of Industrial Ecology* this journal (Dijkema and Basson (2009), Dijkema et al. (2015); see the respective review articles Wood and Lenzen (2009) and Meerow and Newell (2015)). Although the term "network" is frequently used in a qualitative sense (Heijungs 2015), graph theory is a rich source of concepts for quantitative analysis of network structures and parallels to analytical approaches in economics and in ecology have been revealed (Suh 2005). Many graph-theoretical applications draw from theoretical ecology (Odum 1994; Ulanowicz 1997) and analogies between ecosystems and social, economic and industrial systems have been identified (Côté and Hall 1995; Graedel 1996; Korhonen 2001; Bailey et al. 2004). As reviewed in Schiller et al. (2014), graph-theoretical network measures have been applied for describing system structures in Industrial Ecology and for comparing different systems to one another (for a recent application, see Nuss et al. (2016)). Despite the increasing use for analysis of non-trivial structures, graph-theoretical network measures such as "connectedness", "clustering" or "cyclicity" have in few studies been specifically interpreted as relative complexity measures, for example regarding life cycle inventories (Navarrete-Gutiérrez et al. 2015) or industrial ecosystems (Layton et al. 2016). In MFA, complexity measures have not been specifically addressed so far, although graph-theoretical complexity measures are also applicable in MFA. When regarding complexity in the sense of "static complexity", which refers to the "number of parts and their linkages" (Allenby 2009), it appears to be useful to express complexity not as a merely relative, but as an absolute measure, as systems (also material flow systems) may not only differ in their complexity because of different linkage patterns, but also because of varying system sizes. An alternative for analysis of network complexity is presented in this paragraph. It is elaborated specifically for MFA systems and is, as other approaches to complexity in Industrial Ecology, inspired by theoretical ecology.

As material flow systems today cover increasing numbers of materials and regions, there is an interest in identifying similarities and differences of MFA systems (Klinglmair et al. 2016). It has

been observed in different fields of Industrial Ecology, that differences in system structures are to a varying degree to be attributed to actual differences in physical systems, but also to priorities of modelers, the chosen level of detail and the structure of available data (Heijungs 2015). Until now, both comparisons of MFA system structures and evaluations of the impact of MFA input data on MFA results are often limited to qualitative considerations (Klinglmair et al. 2016). A set of measures for specifying and comparing MFA systems by quantitative rather than sheer qualitative means appears to be helpful to facilitate further comparisons of MFA systems and also for communication of MFA results. In this paragraph, such measures are proposed. With a focus on flows, the uncertainty and complexity of MFA systems are formalized as properties of MFA systems and quantitatively expressed in the same, abstract dimension. The information content of MFA systems can naturally be derived from their uncertainty once this uncertainty is quantified. The formal framework used for computation is borrowed from the field of theoretical ecology, as introduced in the following.

2.3.3 Information measures in theoretical ecology

In theoretical ecology, a concept for the description of networks has been elaborated based on work of Rutledge et al. (1976). The concept is constructed around the narrative that the functioning of ecosystems can be understood by means of information theory as a function of system size, proportions of flows in relation to other flows and system structure. Based on these ideas, Ulanowicz (1980) developed a set of aggregate measures for describing the state of ecosystems. The starting point is a perspective on an ensemble of flows, which is, in the notation of Ulanowicz et al. (2009), expressed as

$$H = -k \sum_{ij} \frac{T_{ij}}{T_{..}} \, log \, \frac{T_{ij}}{T_{..}} \qquad \text{Eq. 8}$$

where T_{ij} refers to a flow from agent i to j, $T_{..}$ (a dot means summation over an index) refers to the aggregate of all flows in the system and k is to a positive scaling constant. The measure is used to refer to “system diversity” (Rutledge et al. 1976) or “system capacity” (Ulanowicz 1997). Rutledge et al. (1976) argue that H can be decomposed into two components based on information on each flows’ position in the system configuration, so that, in the notation brought forward by Ulanowicz et al. (2009),

$$X = k \sum_{ij} \frac{T_{ij}}{T_{..}} \, log \, \frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \qquad \text{Eq. 9}$$

and

$$\psi = -k \sum_{ij} \frac{T_{ij}}{T_{..}} \; log \; \frac{T_{ij}^2}{T_{i.} T_{.j}} \qquad \text{Eq. 10}$$

where $T_{i.}$ refers to the aggregate quantity that leaves *i*, $T_{.j}$ refers to the aggregate quantity that enters *j* and T_{ij} refers to the quantity that both leaves *i* and enters *j*. *X* and ψ relate to the information-theoretical concepts of mutual information and conditional entropy, which are measures of association quantifying the relation between two variables. Eq. 9 and Eq. 10 express the dependence between agents *i* and *j* as a function of the flow of matter from *i* to *j* in relation to the aggregate quantities leaving *i* and entering *j*. High *X* specifies that *j* depends predominantly on *i*, high ψ specifies that *j* depends little on *i* but mainly on other adjacent *i*'s. The measures have been applied mainly to ecosystems, for example for examining the functioning of food webs (Wulff et al. 1989; Baird and Ulanowicz 1989) or for comparison of ecosystems (Christian et al. 2005). It has also gained interest in other system-oriented fields such as economics (Goerner et al. 2009) and Industrial Ecology (Kharrazi et al. 2013), where it has been interpreted as a measure for the sustainability of a network. As highlighted by Kharrazi et al. (2013), it is a useful characteristic of the measures that they allow considering both intensive and extensive system properties, a feature which is utilized later in this work.

The measures for the uncertainty and complexity of MFAs of a given size and structure proposed in this paragraph have a focus on the role of flows. They are based on variables relating to the quality of flow data and to the configuration of flows in the system. Their computation is facilitated by Eq. 8, Eq. 9 and Eq. 10. In order to provide a quantity that holds as a reference against which the uncertainty and complexity of a given MFA system can be measured, Eq. 8 is reformulated to express the "informational" system size *S* of an MFA as a function of the number of flows n_F in a system where, for now, all flows *Fi* have the weight 1, and the aggregate of all flows in the system $\sum_i Fi$ (here: $\sum_i Fi = n_F$) is used as the scaling constant *k*, so that

$$S = -\sum_i Fi \cdot \sum_{(Fi)} \frac{Fi}{\sum_i Fi} \log \frac{Fi}{\sum_i Fi} = -n_F \; log \frac{1}{n_F} \qquad \text{Eq. 11}$$

S is a monotonic increasing function of n_F and for systems with an arbitrary number of flows, it is $\lim_{n_F \to \infty} S = \infty$. A binary logarithm is chosen for computation in this thesis and the resulting quantity is

referred to as "informational units". Each individual flow contributes to the magnitude of *S*, i.e., is a component of the sum, which allows quantitatively specifying the contribution of any individual *Fi* to the aggregate system uncertainty and system complexity, as elaborated in the following.

2.3.4 Uncertainty of material flow systems

A typical procedure for filling a given qualitative MFA system with numbers consists of two steps (Brunner and Rechberger 2016). In the first step, *a priori* data for specification of system variables is collected. This data may be incomplete and inconsistent and therefore, in the second step, is balanced and reconciled in an MFA model. This second step increases the completeness and decreases the inconsistencies of data in the model. Ideally, such balanced MFAs provide reliable information on material flow systems. If all flows in a system were known with absolute certainty, their "information content" would be maximal. It would increase with the level of detail of a given system, that is, with the number of flows that are distinguished and correctly specified. As such ideal cases are unrealistic because of data quality limitations, there typically is a remaining degree of uncertainty in MFA results (Laner et al. 2014). As statistical evidence for specification of data quality is frequently limited in MFA, it can be expressed by reliability indicators such as the information defects introduced earlier. Because of data quality limitations, the actual information content of given MFA systems usually is lower than their potential information content, as to the uncertainty remaining in the system. A formal way to express this limitation by quantitative means is proposed in the following paragraphs.

2.3.4.1 Uncertainty of systems with *a priori* data

As information can be understood as the absence of uncertainty, and *vice versa*, *S* allows two interpretations. First, if all flows were known, it can be interpreted as the potential information content of a material flow system. Second, it can be interpreted as the uncertainty of a given qualitative material flow system, where none of the flows is known. This uncertainty may be reduced by integrating data on these unknown flows into the system. In other words, the uncertainty of a system with *a priori* data on flows (U_{ap}) can be understood as a composite of *S*, which reflects the uncertainty of a system without data. Instead of assigning the equal weight 1 to all flows, as it is in Eq. 11, the flows can be weighted by their information defects ID_{Fi}. The uncertainty of a system with *a priori* information (U_{ap}) can then be formulated as

$$U_{ap} = -\sum_{i} Fi \cdot \sum_{i=1}^{n_F} \frac{ID_{Fi}}{\sum_i Fi} \; log \frac{ID_{Fi}}{\sum_i Fi} = -\sum_{i=1}^{n_F} ID_{Fi} \; log \frac{ID_{Fi}}{\sum_i Fi} \qquad \text{Eq. 12}$$

The measure U_{ap} refers to the uncertainty remaining in a quantitative MFA system after data of varying quality on flows is considered. It is $\lim_{ID \to 0} U_{ap} = 0$ and $U_{ap} \leq S$. This becomes clear when considering the simple examples in Table 6 (column "uncertainty") and the case studies in paragraph 2.3.6. The observation that the uncertainty of flows may better be expressed in relation to the uncertainty of other flows in a system (Klinglmair et al. 2016) reflects in Eq. 12, where the uncertainty of each individual flow is not only a function of its specific *ID*, but also expressed in relation to the sum of all *ID* in the system.

2.3.4.2 Uncertainty of balanced material flow systems

By balancing material flow systems, conflicting model input data are reconciled and data gaps are closed. As a result, the uncertainty of a system decreases (i.e., the consistency of a system increases) when it is balanced. Consequently, the uncertainty after balancing (U_b) should be lower than the uncertainty before balancing (U_{ap}). A typical application for balancing material flow systems is the software STAN (www.stan2web.net). In STAN, an algorithm based on the weighted least square method is implemented for data reconciliation. A system with information on flow quantities (linear constraints) is reconciled based on the relation between factors such as standard errors (see Cencic (2016a)). In this work, the information defects are used as factors in data reconciliation with software STAN.

The uncertainty remaining in a system after balancing (U_b) is computed by replacing ID_{Fi} in Eq. 12 by $ID_{Fi,b}$ (information defect of *Fi* after balancing) and it is $U_b \leq U_{ap}$. As both U_b and U_{ap} are composites of *S*, the difference between the actual system uncertainty (U_{ap} or U_b) and the system size *S* is referred to as the information content of a material flow system.

2.3.4.3 Weighted uncertainty of balanced material flow systems

While some flow quantities X_{Fi} are known before balancing (*a priori* data with $ID_{Fi} \epsilon (0,1)$), others are typically unknown (data gaps with ID_{Fi}=1). After balancing a system, all flow quantities $X_{Fi,b}$ in a system are known. Some of these flows may be quantitatively more relevant than others. Intuitively, knowing quantitatively major flows better contributes more to the overall state of knowledge about a material flow system than knowing quantitatively minor flows better. To also consider the quantitative relevance of flows within a system, the uncertainty measure *U* is adapted. Each flow is

weighted by $\frac{X_{Fi,b}}{\sum_i X_{Fi,b}}$, where $X_{Fi,b}$ is the quantity of a balanced flow, multiplied with the number of flows n_F so as not to change the magnitude of the summed *U*-measure. Combining the balanced information defects and the balanced flow quantities, this gives the weighted uncertainty measure $U_{b,w}$, which is

$$U_{b,w} = -\sum_{i=1}^{n_F} \frac{X_{Fi,b} n_F}{\sum_i X_{Fi,b}} \, ID_{Fi,b} \, log \frac{ID_{Fi,b}}{\sum_i X_{Fi,b}} \qquad \text{Eq. 13}$$

By means of Eq. 13, it can be expressed that, given data of good quality on the quantitatively most relevant flows of a system (large $X_{Fi,b}$), the information content of the system is high or, conversely, the uncertainty of this system is low.

2.3.5 Complexity of material flow systems

As motivated by Allenby (2009), complexity may be regarded as a system property which involves both the system size and linkage patterns within a system. This relates to the understanding of Rutledge and colleagues, where a system is maximally complex (or non-trivial) if it consists of many elements and when each of these elements is connected to every other element in the system. Respectively, a trivial network structure, such as a line network of arbitrary length, is of little or no complexity, although it may be of considerable size. Recalling the useful feature of Eq. 9 and Eq. 10 to allow for combined consideration of intensive (system structure) and extensive (system size) dimensions motivates to apply the metrics for aggregated characterization of MFA networks.

Each flow *Fi* is considered to define a subset. In an MFA system, *Fi* connects a source process y_i to a target process z_i. At both its source and target process, *Fi* probably has a number of neighboring flows in the sense of flows originating also from y_i or also entering z_i (see Figure 15).

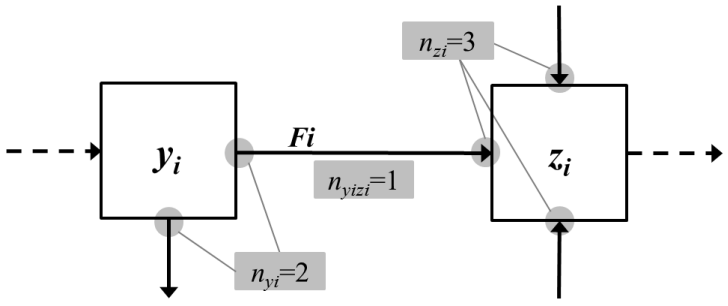


Figure 15: Each flow Fi defines a subset with two characteristics: The outdegree of its source process y_i (n_{yi}) and the indegree of its target process z_i (n_{zi}). A flow from z_i to y_i is referred to as n_{yizi}=1.

Referring to the number of flows leaving y_i and entering z_i as the outdegree of process y_i (n_{yi}) and the indegree of process z_i (n_{zi}), denoting a flow Fi from y_i to z_i as n_{yizi}=1 and considering the total number of flows n_F in a system, Eq. 9 is reformulated as a measure for the triviality T of a system, so that

$$T = n_F \sum_{i=1}^{n_F} \frac{n_{yizi}}{n_F} log \frac{n_{yizi} n_F}{n_{yi} n_{zi}} = \sum_{i=1}^{n_F} \log \frac{n_F}{n_{yi} n_{zi}} \qquad \text{Eq. 14}$$

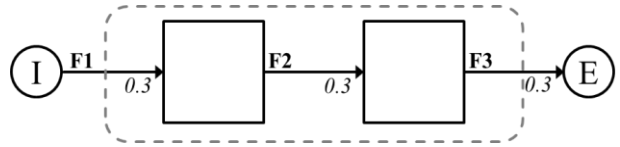
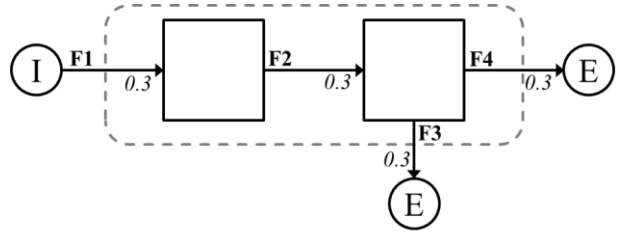
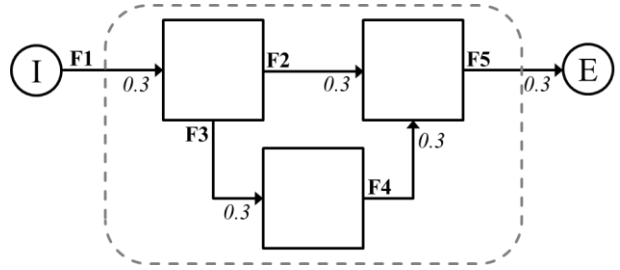
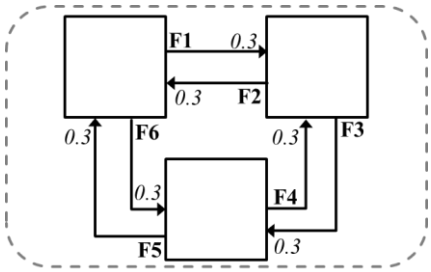
and its counterpart, the complexity C of a system, is formulated as

$$C = -n_F \sum_{i=1}^{n_F} \frac{n_{yizi}}{n_F} log \frac{n_{yizi}^2}{n_{yi} n_{zi}} = -\sum_{i=1}^{n_F} \log \frac{1}{n_{yi} n_{zi}} \qquad \text{Eq. 15}$$

Simple examples are provided in Table 6. In the most trivial topology (a line network, example A), it is S=T and C=0. C increases with n_F and more complicated linkage patterns in example B and example C. Under the condition that no process connects to any other process by more than one flow, systems with n_p processes are maximally complex if they have $n_{F,max}$=(n_p-1)·n_p flows and if each process connects to every other process in the system (example D). In such maximally complex topologies, there always is a T component and, with increasing $n_{F,max}$, it is $\frac{C}{S} \rightarrow 1$ and $\frac{T}{S} \rightarrow 0$. For all above described topologies, it is S=T+C.

Real-world MFAs typically are between the extreme cases illustrated in example A and example D in Table 6. As line networks are untypical topologies of material flow systems, there always is a C component in a realistic MFA. Because of MFA-specific structural limitations, C is never maximal. These limitations include that flows crossing the system boundary originate from or enter processes with an outdegree (in the case of import flows) or an indegree (in the case of export flows) of one. Also, pairs of processes are typically not connected in both directions but in one direction only. According to Eq. 11 - Eq. 15, each individual flow Fi contributes to the aggregated measures by a specific degree, which enables distinguishing flows from one another according to their respective uncertainty in the system context and their configuration in the system structure. This is further elaborated in two case studies presented in the following.

Table 6: Four examples (A-D) for illustration of the proposed system measures S (system size), U (uncertainty), T (triviality) and C (complexity). F1-F5 are the flow numbers. The numbers next to the flows designate the information defects ID_{Fi} (here considered equal for all flows to illustrate the influence of system size on the U measures). The dotted line represents the system boundary, flows crossing the system boundary are referred to as import or export flows.

	Graph	System measures
A		S = 4.8 U = 1.4 T = 4.8 C = 0.0
B		S = 8.0 U = 2.4 T = 6.0 C = 2.0
C		S = 11.6 U = 3.5 T = 7.6 C = 4.0
D		S = 15.6 U = 4.7 T = 3.6 C = 12.0

2.3.6 Application of the measures

The application of the proposed measures has been illustrated in simple hypothetical examples in Table 6. The measures can also be applied to full-scale MFAs, as presented in the following.

Buchner et al. (2014) provide a detailed analysis of aluminum (Al) flows in Austria for the year 2010 (Figure 16). The aluminum MFA consists of 77 flows (n_F=77). The sum of all $X_{Fi,b}$ in the system is about 4600 kilotonnes per year.

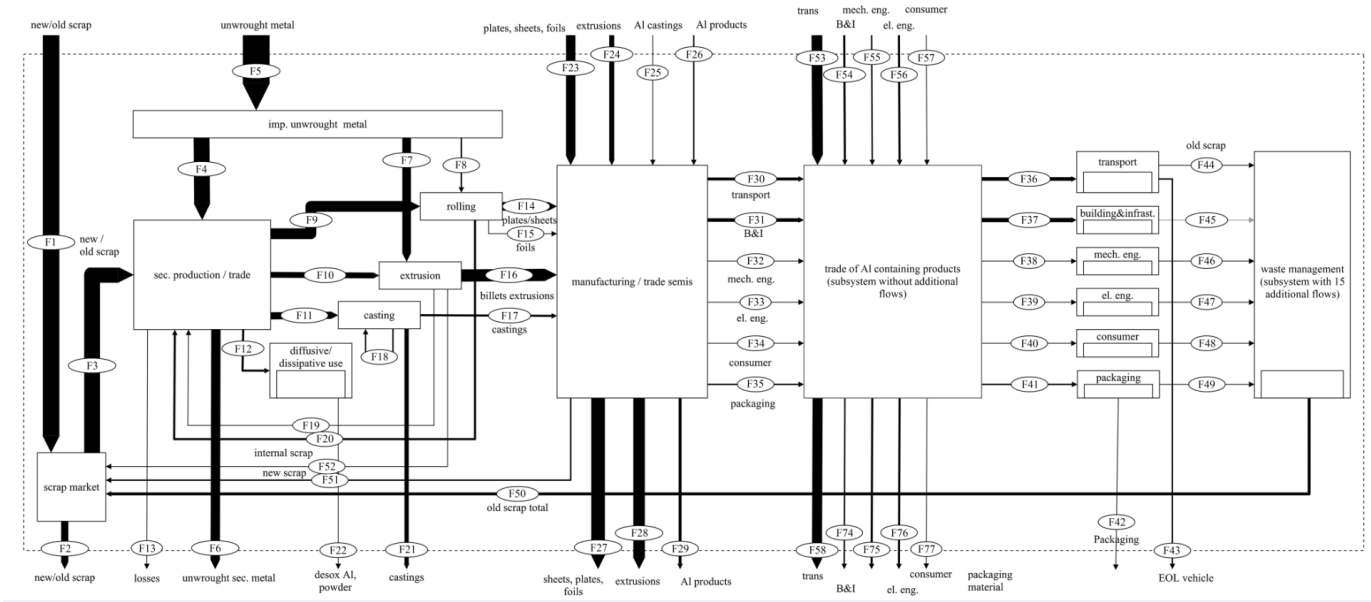


Figure 16: Flowchart of the 2010 Austrian aluminum flow system (Buchner et al. 2014).

In van Eygen et al. (2016) a detailed study of plastics flows in Austria for the year 2010 is presented (Figure 17). The plastics MFA consists of 88 flows (n_F=88). The sum of all $X_{Fi,b}$ in the system is about 15,000 kilotonnes per year.

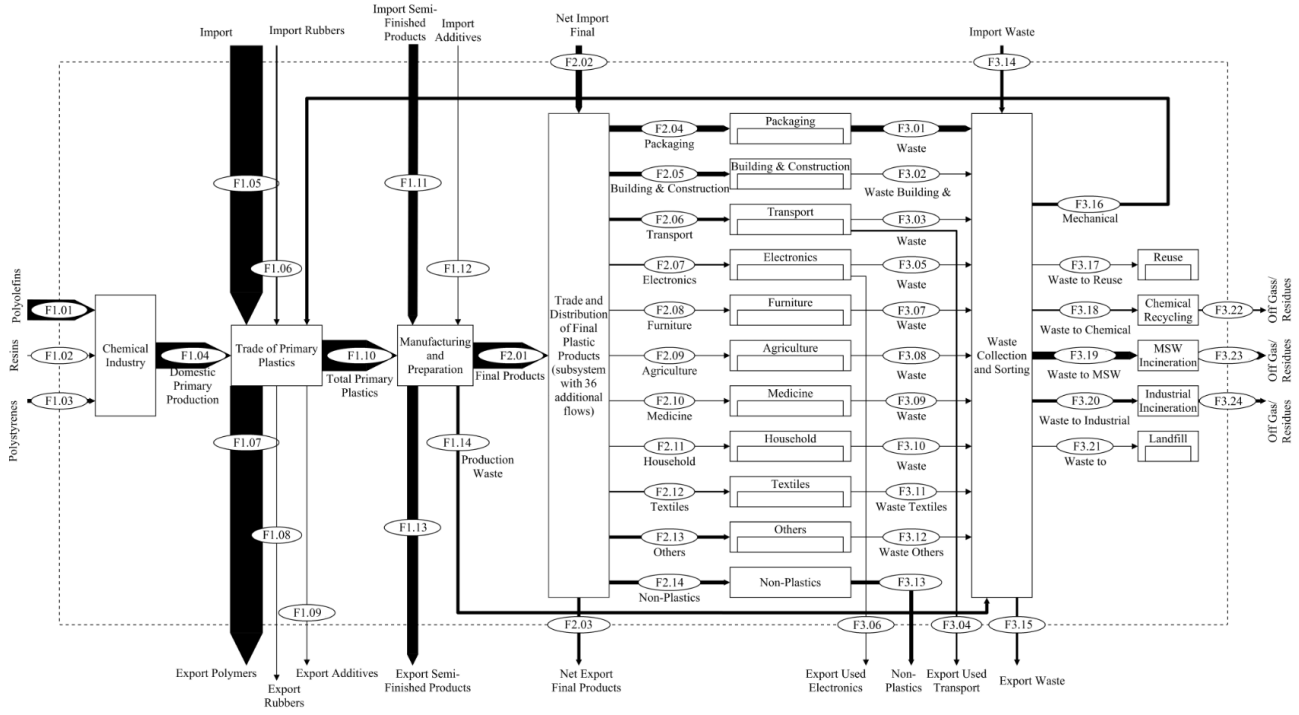


Figure 17: Flowchart of the 2010 Austrian plastics flow system (van Eygen at al. 2016).

The quality of model input data (information defects ID_{Fi}) of the two studies (Figure 16, Figure 17) has been evaluated according to Schwab et al. (2016b). The measures for uncertainty and complexity in MFA (calculated according to Eq. 11- Eq. 15) of the case studies are listed in Table 7.

Table 7: Measures for the information content and structure of the aluminum and plastics systems

	Aluminum	Plastics
System size		
S	483	568
Uncertainty		
U_{ap}	196	335
U_b	99	202
$U_{b.w}$	71	168
Structure		
T	270	296
C	212	272

The absolute magnitude of the measures enables comparing the aluminum and the plastics systems to one another in an absolute sense. The fact that the Al system consists of fewer flows than the plastics system reflects in the measure S (Table 7). Combined absolute and relative comparisons of the measures provide useful information about MFA systems, as illustrated in Figure 18. A list of all input variables and of each individual flows' contribution to the total system uncertainty and complexity is provided in appendix 1.

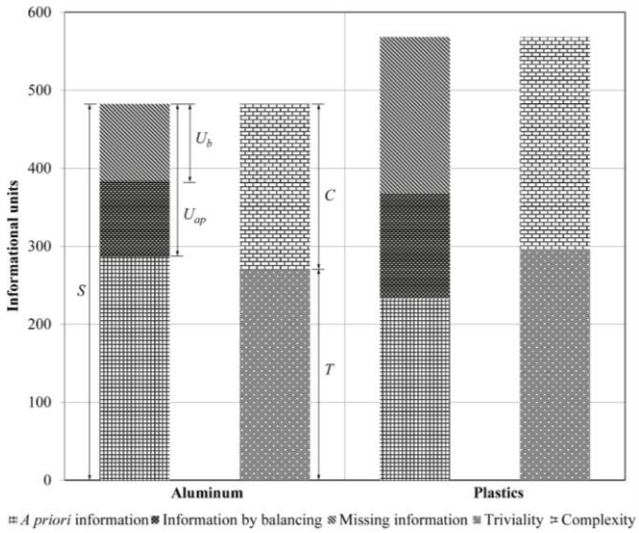


Figure 18: Information content of the aluminum and plastics MFAs (calculated as differences between S, U_{ap} and U_b) and their triviality and complexity.

By means of the ratio of the uncertainty measures *S*, U_{ap} and U_b, the information gain from a qualitative system to a system with *a priori* information and a balanced system can be quantified. The initial uncertainty of the qualitative Al system is *S*=483. Considering the available *a priori* information on flows *Fi*, the uncertainty decreases to U_{ap}= 196. By balancing (data reconciliation and bridging of data gaps), the uncertainty in the system decreases to U_b= 99. In the plastics system, the uncertainty decreases from *S*=568 to U_{ap}=335 and to U_b=202. The plastics flow system is both relatively and absolutely speaking more complex than the aluminum flow system, as it can be seen when comparing the absolute magnitude of *C* and the relation of *C* to *S* of the both case studies. In both systems, $U_{b,w}$ is lower than U_b. This indicates that, both in the Al and plastics system, quantitatively dominating flows are known better than quantitatively minor flows (Table 7). The *a priori* data of the Al system is better than the *a priori* data of the plastics system. This is indicated by the fact that U_{ap} equals less than half of *S* in the Al system, while U_{ap} of the plastics system is only two fifth lower than *S*. By balancing, the uncertainty of the aluminum system decreases by 21% and the uncertainty of the plastics system by 24%. The relation of $U_{b,w}$ to *S* indicates that the aluminum system is known to an extent of 85% and the plastics system is known to an extent of 70%.

2.3.7 Usefulness and limitations

A convenient feature of the uncertainty and complexity measures presented in this thesis is that both phenomena are quantified as formally linked measures and expressed in the same abstract dimension. They enable evaluating the system structure and the state of knowledge on different MFAs or on MFAs at different points in their development process by quantitative means. The information gained by performing MFA procedures can be quantified and compared. As shown in the aluminum and plastics case studies, the information content of material flow systems can instantly be derived from the uncertainty of systems once this uncertainty is quantified. A practical characteristic of the measures *S* and *U* is that they represent both the information needed for construction of a qualitative system of flows and the data for the quantification of these flows. The more flows there are in a system, the higher is the total system uncertainty (resp., the potential information content of a system) and the more and better data are needed to minimize *U*. The presented procedure focuses on flows. Transfer coefficients, stocks and stock change rates, which are also entities that introduce information or uncertainty into material flow systems and that add to the complexity of systems, are not considered. It thus may be useful to implement these entities into the measures in further research. Stock change rates, for example, can be treated identically to import or export flows (which

originate from or enter processes with outdegree or indegree one). That way, they would contribute to *S*, reflect in the structural measures *T* and *C* and, after assigning an *ID*, also in the uncertainty measures *U*.

It has to be recalled that the actual design of a system may for various reasons not be complete and representative, for example because it depends on probably arbitrary decisions of agents (Heijungs 2015). Although differences in system design have been identified as to be relevant in MFA, distinctions and comparisons regarding system structures are intricate (Klinglmair et al. 2016). The proposed measures *T* and *C* provide an improved basis for future analyses and comparisons. While the term "complexity" of MFAs has been used in a qualitative manner (Klinglmair et al. 2016), it can now be expressed by quantitative means. Increasing complexity *C* of a system relates to increasing information demand of this system. Practically speaking, this means that an agents' effort for data acquisition increases with *C*. Despite these possibilities, the proposed procedure for structural analysis is limited regarding adequate representation of some MFA-untypical topologies, which may be chosen by agents performing an analysis. This applies for example to loops (flows leaving and entering the same process in the studied unit time interval), such as flow 18 in the aluminum case study, and to parallel flows (flows from and to the same source and target process), such as flows 14 and 15 in the aluminum case study (Figure 16). Parallel flows result from flows being represented in a more disaggregated manner than processes and can in virtually all cases be prevented by respective system design, that is, by disaggregating either the source process or the target process, for example by use of subsystems (as it is for flows 30-35 in Figure 16). For one topological particularity, the characteristic of the structural measures that the contribution of each *Fi* to *C* and *T* is expressed in relation to the total number of flows in the system yields notable, though unproblematic, results. For a *Fi* that connects processes with particularly high outdegree n_{yi} and indegree n_{zi}, so that $n_{yi} \cdot n_{zi} > n_F$, the contribution of this *Fi* to *T* takes negative values (cf. the concept of "pointwise mutual information" in information theory). This is not the case for any flow in the two systems analyzed in this thesis and, though conceptually possible, improbable in real-world MFAs. Such situations reflect particularly high structural heterogeneity of systems, which entails that the proportions different flows have in *C* and *T* shift. A *Fi* with a negative contribution to *T* has a proportionately higher contribution to *C* and the aggregate measures *C* and *T* sum up to *S*. The usefulness of the proposed measures for analysis and comparison of MFA systems is further illustrated in the following chapter.

3 Cases

For illustration of the concepts introduced in chapter 2, the four Austrian material flow systems introduced in the previous paragraphs are analyzed and compared: phosphorus (Figure 4), palladium (Figure 11), plastics (Figure 17), and aluminum (Figure 16). The phosphorus system (Zoboli et al. 2015) has been addressed in paragraph 2.1, the 2011 palladium (Laner et al. 2015a) in paragraph 2.2, and both the 2010 plastics system (van Eygen et al. 2016) and the 2010 aluminum system (Buchner et al. 2014) in paragraph 2.3. In paragraph 2.1, only the main system on phosphorus has been studied (n_F=72). In this chapter, the whole system including subsystems is analyzed (n_F=122). Note that identifying differences in the studies is not to say that one MFA is done "better" than another (MFA results depend on many factors, such as the specific questions authors intend to answer and the availability of both qualitative and quantitative information), but to show that the proposed procedures can be used to detect and quantify differences in MFAs.

In paragraph 3.1, the databases of the four case studies are analyzed and differences or similarities are described by comparison of selected data attributes. In paragraph 3.2, information defects (*ID*) of the four case studies are quantified and compared. In paragraph 3.3, the structure of the systems is analyzed and the information content of the material flow systems is quantified and compared.

3.1 Data characterization

The data structures of material flow systems vary with scope and goal of the studies, and consequently also reflect in the comparisons provided in this paragraph. Comparisons reveal interesting differences and similarities of material flow systems, as reflected in their database. Before moving to the data attributes, general characteristics of the four compared systems and their databases are compared in Table 8. Phosphorus is the most detailed out of the four studied systems, as it also reflects in the quantity of collected data elements and the number of entities (such as rates and concentrations) per study. The share of data gaps is highest in the palladium and the plastics studies, where more than one third of the flows is unknown *a priori*, and least for the phosphorus study, where 2% of the flows are unknown *a priori*. More details on the databases are provided in Appendix 1.

Table 8: Characteristics of the four studied material flow systems and their databases

Database characteristics	Quantity			
	Phosphorus	Palladium	Plastics	Aluminum
Number of flows n_F	122	25	88	77
Unknown flows (*a priori*)	3	9	33	18
Number of processes	56	12	37	28
Number of subsystems	8	2	1	1
Number of stocks	8	5	12	8
Number of data elements	399	48	70	73
Number of entities	258	30	70	73
Average number of entities per flow	2.2	3.2	1.3	1.2
Average number of data elements per entity	1.5	1.6	1	1
Share of flows that are described directly by autonomous data (%)	5	10	63	63
Isolated values (%)	78	67	100	100

The databases of the plastics and aluminum study are of similar size. For both studies, comparatively high shares (63%) of directly applicable data (tonnes of plastics, resp. aluminum, per year) are available in the form of isolated values. For phosphorus and palladium, flows are typically quantified based on information on material quantities and substance concentrations, and autonomous data are available for few flows (5%, resp. 10%). Information on concentrations is frequently based on more than one reference or sample, which reflects in the lower share of isolated values in the phosphorus and palladium studies. Differences in the databases become also evident when comparing the producer categories, that is, the shares of data elements provided by authorities, by science or economy, or by civil society actors (Figure 19).

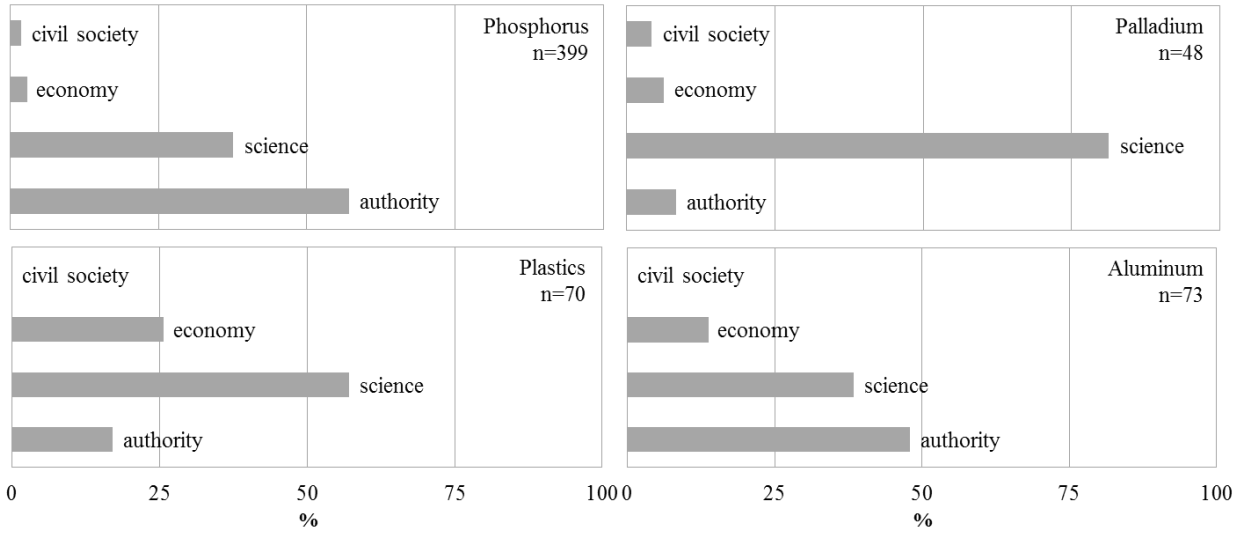


Figure 19: Producer category of data elements.

Generally speaking, most data comes, in varying shares, from authorities and science. Civil society actors, such as consumer associations or environmental organizations, play a subordinate role as information producer in the four studies. Shares of data from economy vary, presumably influenced by factors such as economic relevance of the material, market size (small markets are observed to rather keep their data confidential), and absence of data from authorities in the form of official statistics, what makes the focus of MFA practitioners shift towards agents from economy as information sources. In case there are data neither from economy nor from authorities and, at the same time, empirical scientific information is short, MFA modelers increasingly rely on data derived from assumptions or from speculation, as it is the case for palladium in Figure 20 (note: information based on reporting, such as data from official statistics, is also considered "empirical" here, since the data collection processes behind these statistics ideally are empirical).

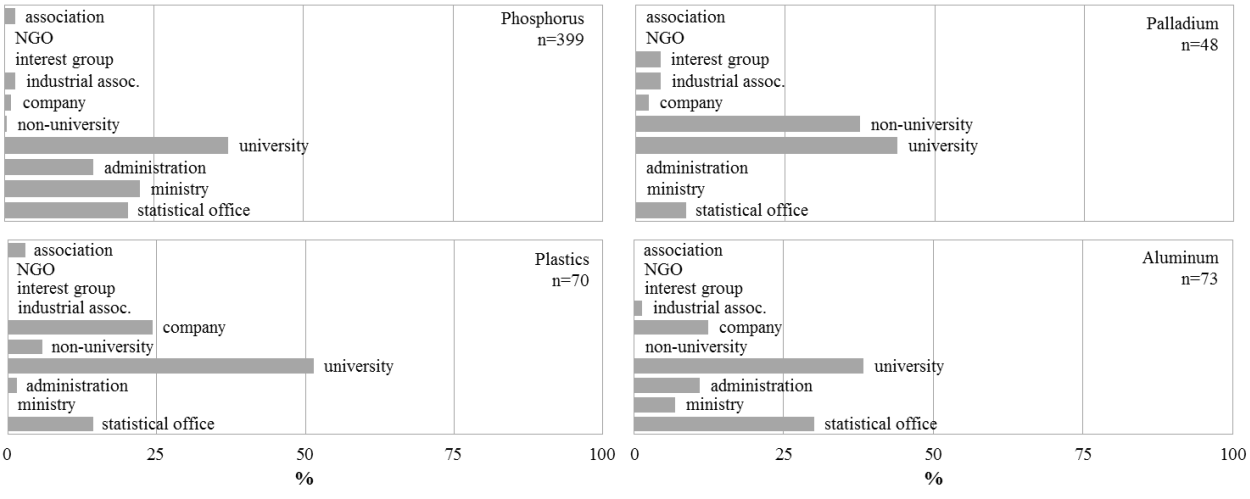


Figure 20: Origination category of data elements.

Figure 20 illustrates a central difference between MFAs, which is the share of empirical versus derived information. From phosphorus and aluminum to plastics to palladium, the MFAs are decreasingly based on empirical information. For aluminum, trade statistics explain a major share of the results plotted in Figure 20. Increasing public interest in phosphorus as a pollutant and as a scarce resource reflects in increasing number of science-based reports published by ministries. In the plastics MFA, a vast share of information is based on models and calculations (origination category "derived (mainly from data)"). While Figure 20 illustrates a difference between the MFAs, Figure 21 shows a similarity, which is the format of collected information.

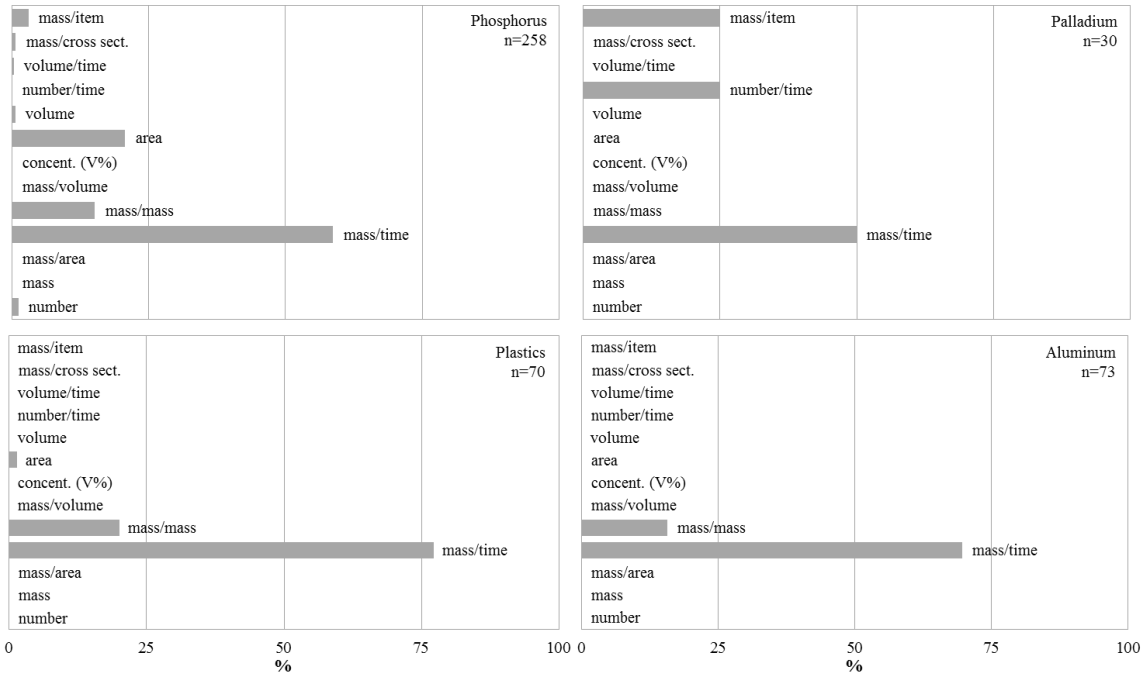


Figure 21: Entity class of information elements.

The entity class is widely homogenous for different MFAs as, obviously, most data are collected in the format mass/time. Yet, the diagrams plotted in Figure 21 show differences between the studies, which are either because of conceptual differences or because of specific material characteristics or applications: For palladium, concentrations have not been considered in the format mass/mass (as for phosphorus and aluminum; note that in the plastics MFA, mass/mass are transfer coefficients), but as concentration per item. For phosphorus, areas are an important additional class of entities, reflecting the relevance of phosphorus in fertilizers. Differences between studied materials reflect, to a higher degree, in Figure 22, where information elements of the four studies are compared regarding the type of good they refer to.

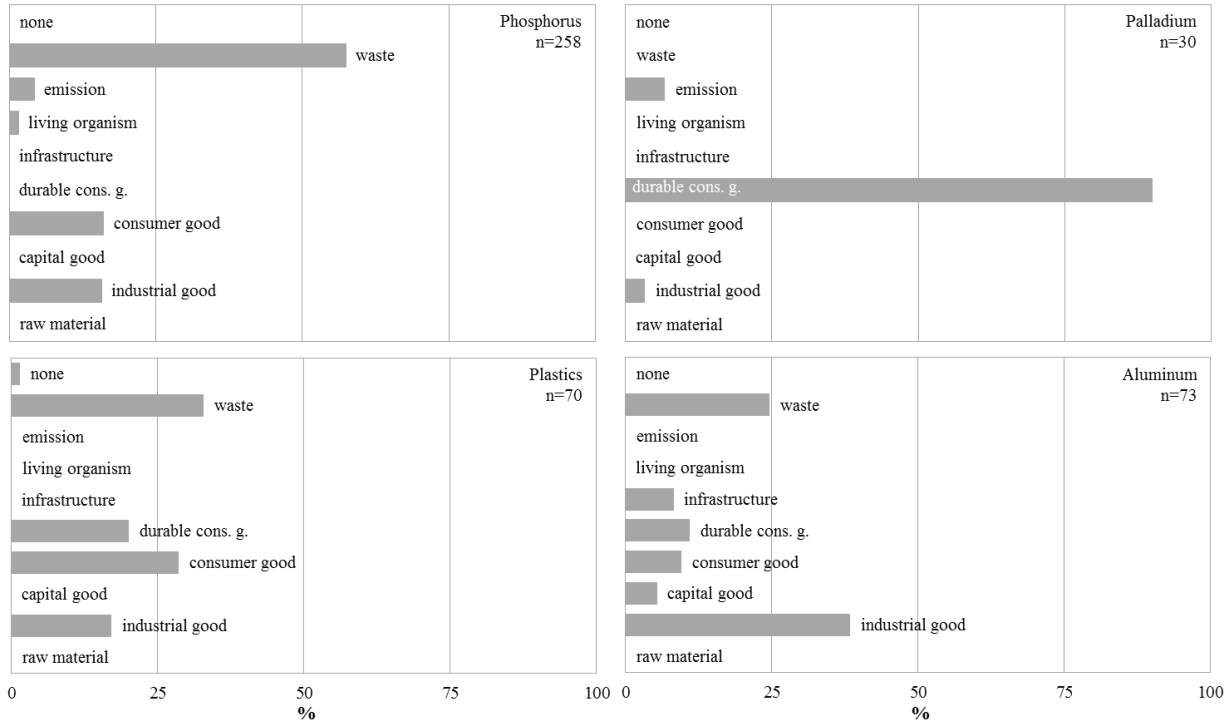


Figure 22: Type of good of the entities.

Most information collected for the phosphorus study relates to waste materials (such as sewage, manure). In the palladium case study, data has mainly been collected for palladium in automobile catalysts (durable consumer goods). Data on plastics has predominantly been collected on consumer goods and waste. The data structure of aluminum indicates a broad range of relevant product types, where industrial goods (such as semis and ingots) were the most data-intensive. The aluminum study has a focus on understanding aluminum recycling, which reflects in the share of data on waste (end-of-life materials such as scrap have here been categorized as waste).

According to the share of collected information elements, material flow systems are dominated by different natural or anthropogenic processes (Figure 23). As reflected in the data structure, market activities influence all four analyzed material flow systems. While processes of the natural environment (biosphere, geosphere) impact on the phosphorus system, technology appears to be, as inferred from the information base, a more relevant driver of the palladium, plastics and aluminum systems.

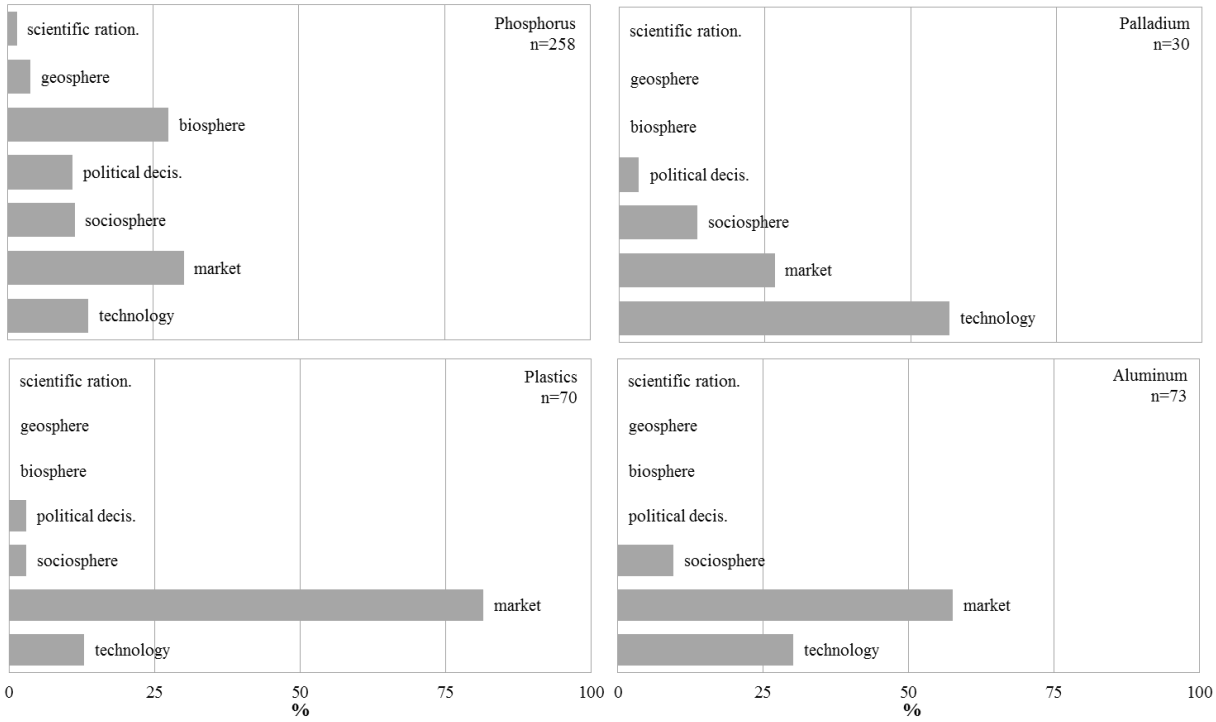


Figure 23: Primary determination of entities.

The data characterization matrix (paragraph 2.1.2) consists of more than the here compared five data attributes. Information for further comparisons is provided in Appendix 1. Data quality-relevant attributes are used for quantification of information defects in the next paragraph.

3.2 Data quality evaluation

The results of the data quality evaluation procedure proposed in paragraph 2.2 are displayed in Figure 24 (phosphorus), Figure 25 (palladium), Figure 26 (plastics) and Figure 27 (aluminum). Documentations of data attributes, ID_i and ID_{tot} are provided in Appendix 1. Note that here, the quality of *a priori* data is evaluated, and that low information defects refer to good data quality.

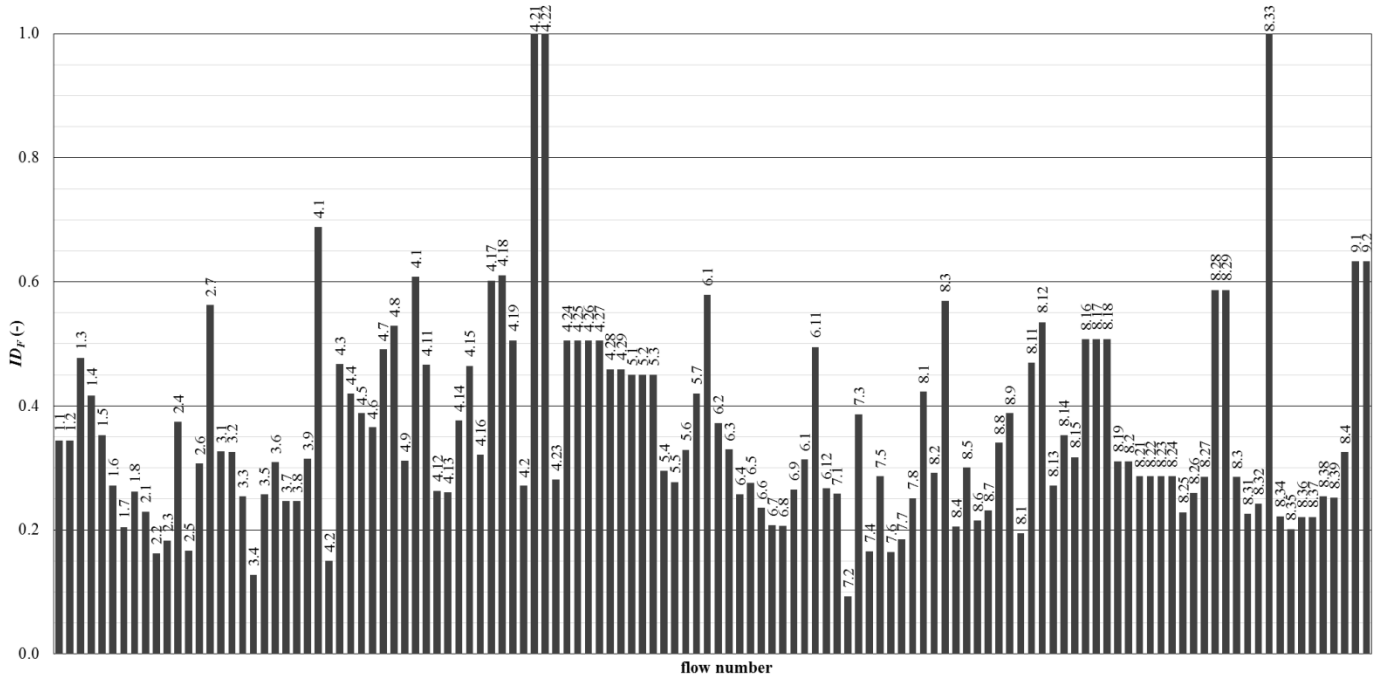


Figure 24: Information defects (ID_F) of the phosphorus case study.

There are few data gaps in the phosphorus case study (Figure 24). Most ID_F are between 0.2 and 0.5.

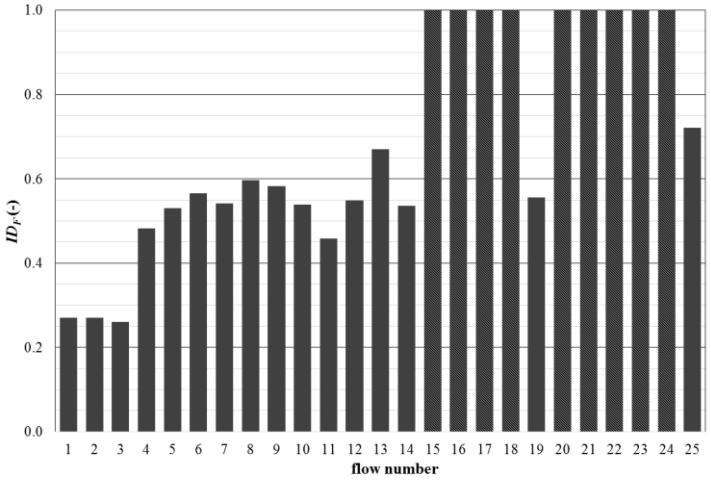


Figure 25: Information defects (ID_F) of the palladium case study.

In the palladium study (Figure 25), ID_F increase and the number of data gaps increases when moving downstream. ID_F are mostly between 0.2 and 0.6.

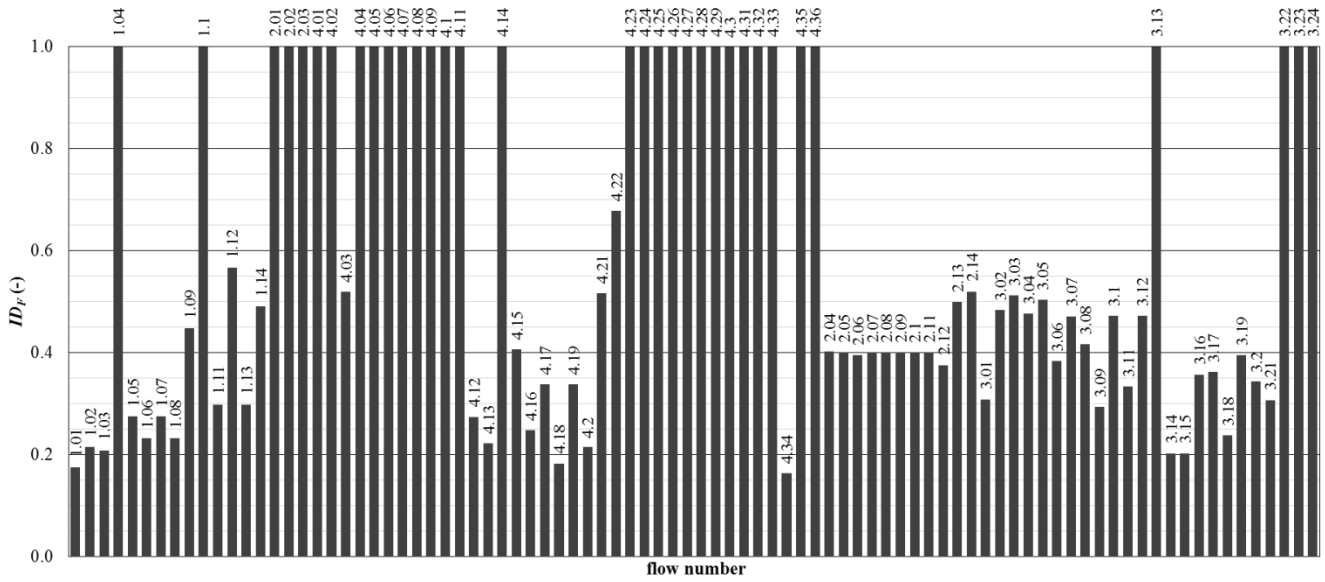


Figure 26: Information defects (ID_F) of the plastics case study.

In the plastics system, ID_F are mostly between 0.2 and 0.5. Data gaps are predominantly in the use sector (flow numbers 4.1-4.36).

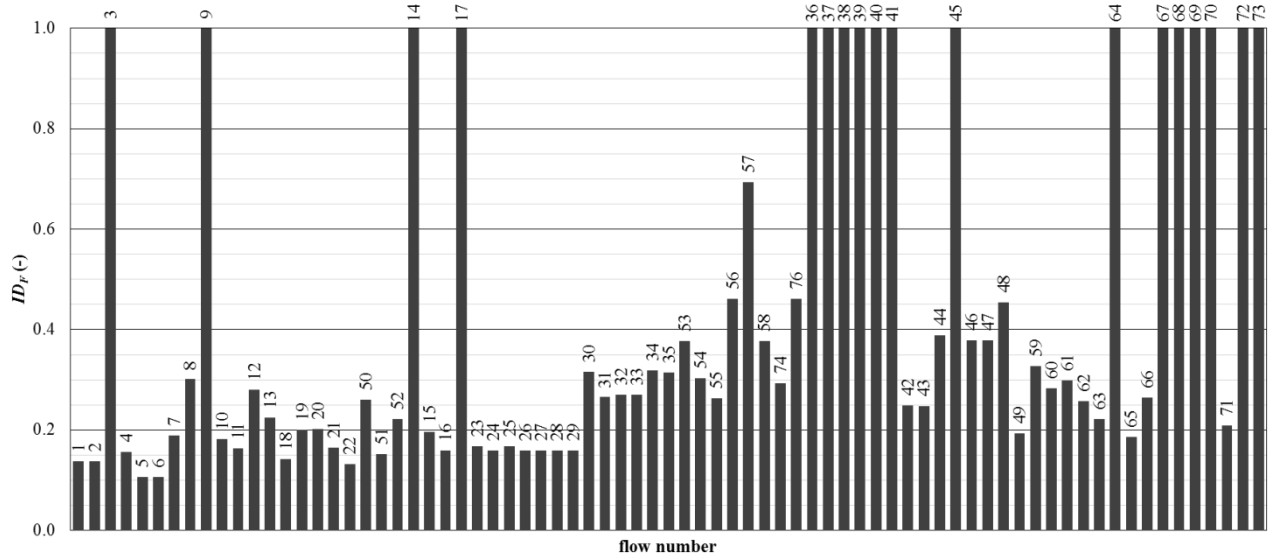


Figure 27: Information defects (ID_F) of the aluminum case study.

The ID_F in the aluminum study are comparatively low (mostly between 0.1 and 0.4), but there are more data gaps than in the phosphorus case. In Figure 27, a slight trend of increasing ID_F when moving from primary production and manufacturing towards use and end-of-life sectors can be observed. Most data gaps are in the use (flows 36-41) and in the waste management (flows 64-73) sectors.

The differences in the ID_F patterns and the share of data gaps in the case studies reflect in Figure 28: Phosphorus, the upper curve, has comparatively low ID_F and few data gaps. Aluminum has, in

comparison to the other studies, the lowest information defects but a number of data gaps. Plastics and palladium have a comparable share of data gaps, but the ID_F of the plastics study are lower.

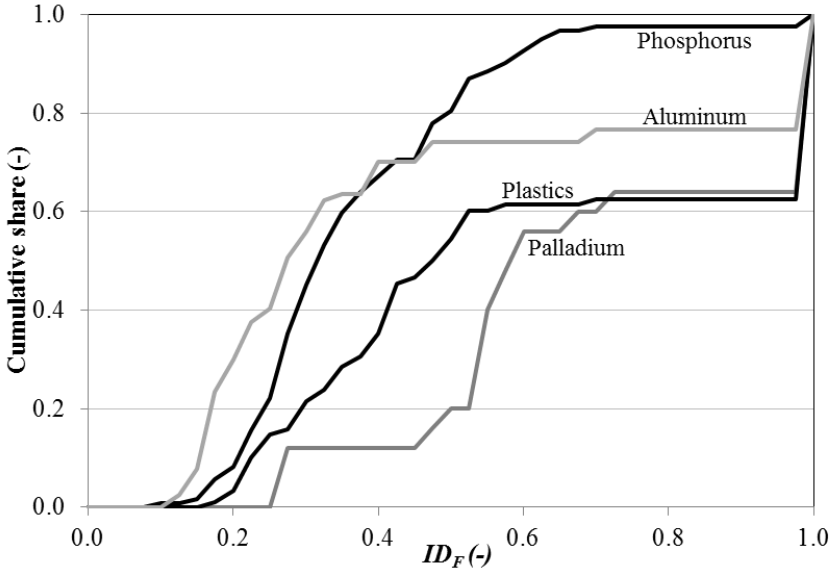


Figure 28: Cumulative distribution of ID_F in the four case studies.

According to the ID_F and with reference to Figure 28, the four studies can be ranked: Regarding the model input information, the phosphorus and the aluminum studies seem to be more reliable than the plastics and palladium case studies. Beyond this relative, normalized comparison, the system metrics proposed in the following paragraph facilitate absolute comparisons of studies by their information content and system structure.

3.3 System structure and information content

An obvious difference between the phosphorus, palladium, plastics and aluminum MFAs is the number of flows n_F in the system (see Table 8 in paragraph 3.1). This reflects in the measure S (Eq. 11), which represents the uncertainty of a qualitative material flow system. Considering the ID_F (paragraph 3.2) in the uncertainty measures (Eq. 12, Eq. 13) and applying the measures for network structure (Eq. 14, Eq. 15) results in metrics useful for distinguishing the four material flow systems (see Table 9).

For every system, the uncertainty decreases from the qualitative system (S) to a system with *a priori* data (U_{ap}) to a balanced and reconciled system (U_b). The measure of uncertainty weighted by flow quantities is designated as $U_{b.w}$.

Table 9: Measures for the information content and system structure of the four case studies, in "informational units" (see paragraph 2.3)

	Phosphorus	Palladium	Plastics	Aluminum
System size				
S	846	116	568	483
Uncertainty				
U_{ap}	300	78	335	196
U_b	229	53	202	99
$U_{b.w}$	234	55	168	71
Structure				
T	475	92	296	270
C	371	24	272	212

The information content of material flow systems is specified as the difference between the uncertainty of a qualitative system (S) and the uncertainty of a quantitative system (U_{ap}, U_b, or $U_{b,w}$), that is, as the amount of uncertainty erased (amount of information gained) during the preparation of an MFA. The uncertainty measures listed in Table 9 are visualized in Figure 29 (left bar for each case study).

The measures T and C reflect the different system structures of the MFAs. Increasing T refers to systems of increasingly trivial structure, increasing C refers to systems of increasingly complex structure. T and C as listed in Table 9 are visualized in Figure 29 (right bar for each case study) and enable distinguishing material flow systems by their system structure. The phosphorus system is – absolutely speaking – the most complex system as it has the highest C. Considering the ratio between C and S, the plastics system is – relatively speaking - more complex than the phosphorus system.

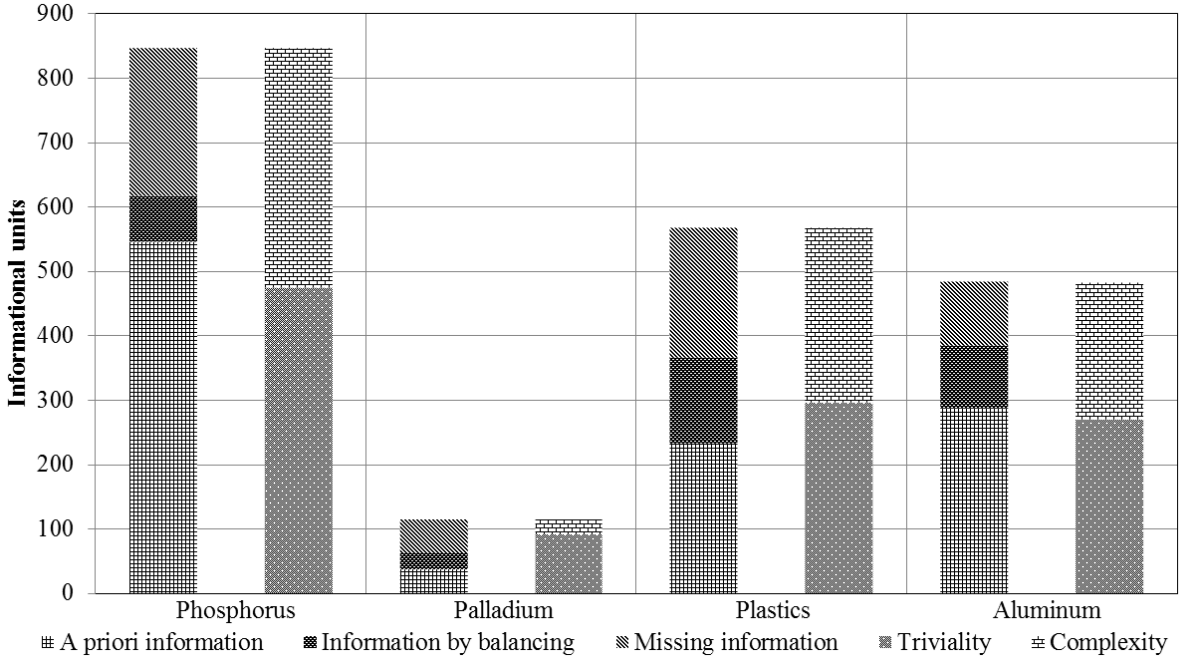


Figure 29: Measures of the information content (left bars) and system structure (right bars) of the four case studies, in "informational units" (see paragraph 2.3).

The results visualized in Figure 29 enable material flow systems to be distinguished by their information content. Comparing the plastics and aluminum cases, it can be seen that, after balancing, the absolute information content of the plastics study is similar to the information content of the aluminum study. However, at the same time, the absolute uncertainty remaining in the plastics system is more than double the uncertainty remaining in the aluminum system. It can also be said that the palladium study has, absolutely speaking, the lowest information content, and, relatively speaking, the highest uncertainty. In relative terms, based on a comparison of S and $U_{b,w}$, it can be said that the phosphorus system is known to the extent of 72 %, the palladium system to an extent of 53%, the plastics system to an extent of 70% and the aluminum system to an extent of 85%. In absolute terms, the phosphorus study has the highest information content.

3.4 Merging the case study results

Combinations of the results of the data characterization (paragraph 3.1), data quality evaluation (paragraph 3.2) and system measures (paragraph 3.3) provide various insights into the informational basis of material flow systems. A few reflections on the results are presented in this paragraph, with a focus on assessing how reliable MFA results are. The usefulness of distinguishing material flow systems by their information content and their system structure is addressed in chapter 4.

In paragraph 3.1, the databases of four MFA systems have been characterized by attributes such as the origination and format of data elements, or the type of real-world objects the data refer to. The results show that the phosphorus and aluminum flow data exhibit a higher percentage of empirical support than the plastics data, which is mainly derived from other data, and than data of the palladium study, which is mainly derived from assumptions. From the comparisons provided in paragraph 3.1, one may infer that, according to the databases, the results of the phosphorus MFA are more reliable than the results of the aluminum MFA, which are more reliable than the results of the plastics MFA and than the results of the palladium MFA.

Beyond comparisons of MFA systems by means of patterns in their databases, the results of paragraph 3.2 enable material flow systems to be distinguished by their data quality and data gaps. The aluminum database consists of better data than the phosphorus database but, at the same time, has a higher share of data gaps. According to paragraph 3.2, the data on both phosphorus and aluminum are of better quality than the data of the plastics database, which also has a much higher share of data gaps. The palladium database has both a high share of data gaps and includes mainly data of poorer quality.

In paragraph 3.3, data quality and data gaps are both integrated into measures of information content. While *a priori* data contribute more to the information content of the phosphorus than of the aluminum study, also because of the lower number of data gaps, relatively more is known about the aluminum system after balancing. For palladium, both the information gained by collection of *a priori* data and by balancing is lower than for all three other studies. While the absolute information content of the phosphorus MFA and the relative information content of the aluminum MFA are the highest, the information gained by system balancing is, absolutely and relative to total system uncertainty, highest in the plastics MFA. Based on the case studies presented in this chapter and according to the procedures proposed in this thesis, it appears that MFA data quality cannot *per se* be judged based on patterns in the data structure alone since data quality depends also on combinations of data attributes and is specific for every flow in its particular application context. In the same sense, the information content of MFAs cannot *per se* be inferred from information on data quality since it also depends also on the configuration of flows in the system context. A combination of the steps, as elaborated in this thesis, however, expresses the phenomena “data characteristics”, “data quality” and “information content” as connected features and provides thorough documentation, reaching from the very components of an MFA, data and its attributes, to aggregate measures for system description.

4 Scientific contribution and limitations

This chapter concludes the thesis with general considerations on the benefits and limitations of the proposed procedures and metrics. For more specific discussions on the data characterization framework, refer to paragraph 2.1.4, for discussions on the data quality evaluation procedure, to paragraph 2.2.4, and for discussions on the measures for information content and system structure, to paragraph 2.3.7.

In this thesis, information in MFA has been framed as an object of research. Limitations to reliability of MFA results have been attributed to problems of limited knowledge such as the absence of statistical information, data quality shortcomings, or the presence of subjectivity. The limitations of MFA information have been conceptualized as a degree of belief and formalized as so-called information defects. As to the premise that every piece of information used for the compilation of an MFA impacts on the uncertainty of the results, these information defects are elaborated on basis of a detailed information inventory. Recognizing uncertainty as a property of systems, the data quality of each flow in a system in combination with information on the system size reflects in measures of the information content of a given MFA. These measures can be used to compare MFAs of different substances, regions or years to one another. For example, phosphorus studies of Austria, Denmark and the Netherlands can be analyzed regarding structural and informational differences. It can be determined, for example, whether MFAs of different industry metals all lack reliable information in similar sectors. By use of the measures for system structure, it can be determined whether structural differences between systems can be attributed to particular sectors, such as the production or the recycling sector.

The value of the uncertainty concept, which is a constituting part of this thesis, is that it does not aim, as other MFA uncertainty concepts (Hedbrant and Sörme 2001; Laner et al. 2015b), to quantify uncertainty ranges, which may convey the impression of empirical evidence even if an MFA is actually not backed by statistical information. In avoiding the use of uncertainty ranges, uncertainty is here quantified in an abstract dimension and material flow systems can be compared both in an absolute and a relative sense. Besides comparison of MFA systems by means of aggregate measures, the procedure proposed in paragraph 2.3 is helpful for identifying weaknesses in MFA systems, that is, for identifying particularly certain or uncertain flows or sectors in a system. This can be represented in a convenient way by means of flowcharts, as proposed in Figure 30. It illustrates, that in the aluminum system (Figure 16), quantitatively major flows contribute most to the system

uncertainty $U_{b,w}$, even though they are mostly known better than quantitatively minor flows (compare Figure 27).

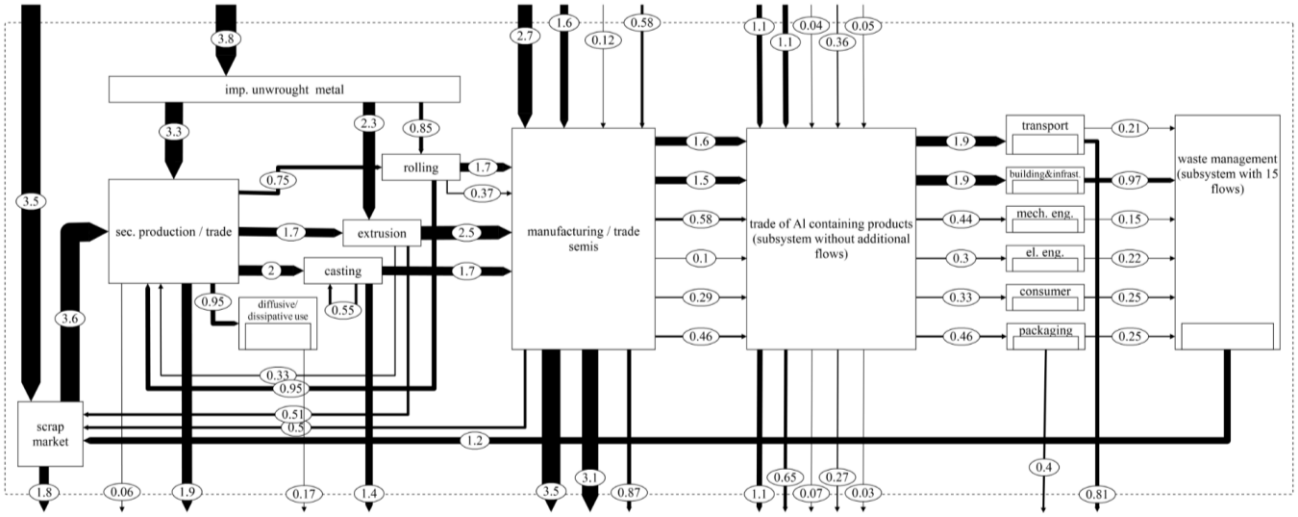


Figure 30: Uncertainty $U_{b,w}$ of the flows in the aluminum MFA displayed as a flowchart. Flow widths are proportional to the weighted uncertainty per flow (Eq. 13). For the total system, it is $U_{b,w}$=71.

The differences between the uncertainty measures are useful for system design as they help detecting potential flaws in the MFA: The difference between U_{ap} and U_b indicates the degree to which *a priori* data of an MFA system is conflicting and has to be manipulated to meet mass balance constraints (data reconciliation). A comparatively high degree of data reconciliation indicates either that the respective material flow data are inconsistent, or that the qualitative system is unrealistic and has to be revised (or both). Additional flowcharts for illustration are provided for the plastics and aluminum case studies in appendix 7. The results for all four case studies are provided in appendix 1.

Representations of MFA systems are not always identical to the physical reality, partially because they are designed based on probably limited quantitative and qualitative information, as argued in Klinglmair et al. (2016). Typically, the level of aggregation or disaggregation in MFA is not only influenced by the scope of a study, but also by information availability. To find the right balance between aggregation (possible loss of information) and disaggregation (possible increase of uncertainty) in view of the available database, the measures S and U_{ap} proposed in paragraph 2.3 can be utilized: The more flows there are in a system, the higher is S and the more and better data are needed to minimize U_{ap}. In an informationally optimized system, it is $(S\text{-}U_{ap})/S \rightarrow 1$. This optimum can be reached by increasing S (distinguishing more flows so that $S \rightarrow \infty$) while decreasing U_{ap} (incorporating better data so that $U_{ap} \rightarrow 0$), or by finding the optimal S for a given U_{ap}. This antagonistic interpretation of S and U_{ap} shows that, depending on the available information basis, it is

usually not helpful to increase the resolution of a system when no adequate data are available. By contrast, given rich and detailed information, increasing the system resolution is beneficial in the sense of increasing *S* and decreasing U_{ap}. When increasing the resolution of a system (when distinguishing more flows in a system), the system complexity *C* increases. Flows that contribute most to *C* and sectors of particularly high or low complexity can be identified and compared quantitatively. As illustrated in Figure 31, the aluminum system as prepared by Buchner et al. (2014) is most complex in the production and trade sectors.

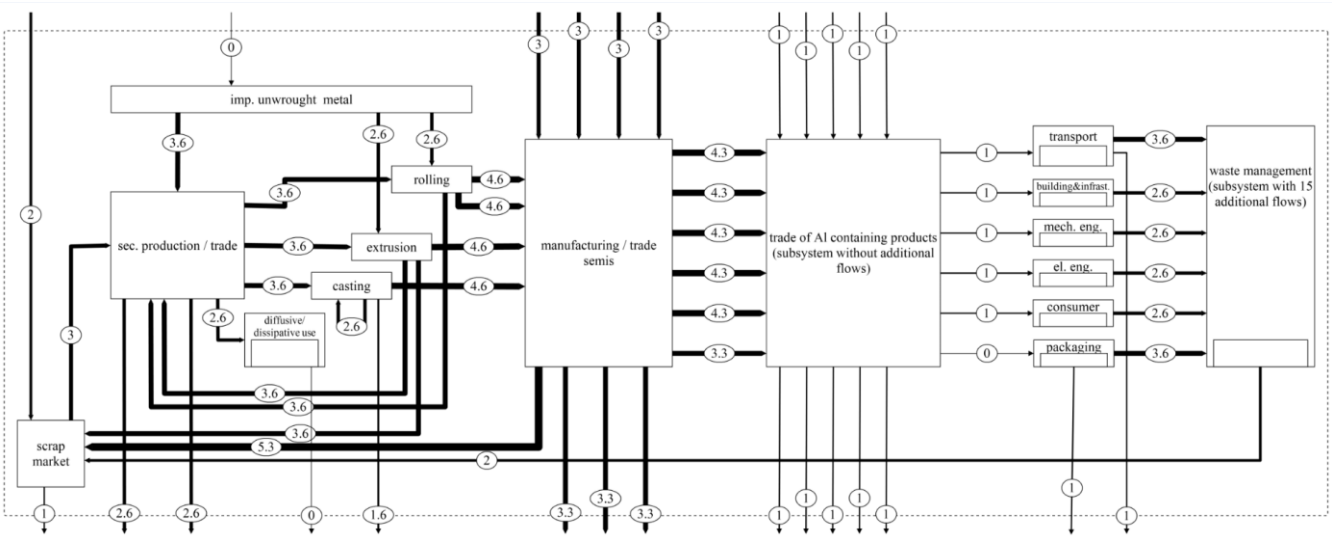


Figure 31: Complexity C of the flows in the aluminum MFA displayed as a flowchart. Flow widths are proportional to the contribution per flow to C (Eq. 15). For the total system, it is C=212.

Beyond the applications of information defects (*ID*, paragraph 2.2) in the measures presented in paragraph 2.3, they can also be understood as new quantities for application as variables in other MFA procedures. For example, *ID* can be applied as variables in MFA models, for example in data reconciliation algorithms that include indicators of data quality (such as Kopec et al. (2015)). *ID* could also be a means for estimating standard deviations, combining aleatory uncertainty and epistemic uncertainty, as proposed by Cencic (2016b). As uncertainty ranges are, despite their limitations, convenient for MFA practitioners, *ID* may also be translated to symmetrical or asymmetrical uncertainty ranges. Here, the data characterization matrix (DCM) can be useful for more information-specific characterization of data uncertainty ranges. Instead of generally assuming normal distributions, different distribution geometries may be systematically assigned to different entities based on attributes documented in the DCM proposed in paragraph 2.1, for example based on the data attribute "property type" (intensive or extensive entities) or other attributes that describe the nature of the entity considered. The formal requirements for computing arbitrary distributions in

MFA have been elaborated (Cencic and Frühwirth 2015). Further probabilistic and possibilistic methods for data manipulation in MFA, as proposed for example in Laner et al. (2015b) and Džubur et al. (under revision), require data quality measures as model input parameters and the *ID* could be applied.

The DCM (paragraph 2.1) has been designed to cover a wide range of data characteristics, more than are actually useful for data quality evaluation. This is to conserve meta-information which may otherwise get lost, and enables interesting evaluations of MFA databases, as illustrated in the case studies in paragraph 3.1. Certainly, the DCM requires rich input of meta-information, and some of this meta-information may not be available or may have been lost over time. In such cases, MFA practitioners could directly estimate the information defects and move on to quantification of information content according to paragraph 2.3. Such estimations may be accurate and useful for further analysis. However, the analysis would then not only miss out the comprehensive documentation of meta-information but, as always when estimations are used without clear documentation, may lose some degree of transparency and replicability.

It is important to note that subjectivity is an inevitable component of the proposed concept. Agents with different perspectives, backgrounds and incentives may have differing perceptions of applied data. This may result in varying data characterization and, consequently, in different *ID* specifications and thus in differing information content. However, it has been argued that subjectivity is a general component of information (Arndt 2004). As subjectivity is unavoidable, the concept presented is designed to restrict subjectivity regarding MFA information by systemizing the characterization and evaluation procedure. Despite the systematic procedure, it should be recalled that the information defects (paragraph 2.2) are, although the formalizations have been carefully elaborated, not statistically backed measures but systematic, formalized estimations. They should be understood as belief indicators. As long as there is no statistically exploitable data, MFA practitioners are limited to using estimations such as the information defects proposed here. Statistical methods should be applied instead whenever the available data are sufficient. It has been widely argued for improvements in data quality and data availability and relevant players have been identified (Wiedmann et al. 2011; Rechberger 2015). One can hope that in the future, available data will increasingly allow use of statistically established methods in MFA data quality evaluation.

Beyond the contributions in a scientific context, the proposed measures bring important benefits for communicating MFA results to third parties such as decision makers in industry and policy making. Lazarevic et al. (2012) reflect on the ability of tools in Industrial Ecology to objectively inform decision makers, that is, to provide an adequate knowledge base for making informed decisions. They argue that value choices, subjectivity and perspectives impact not only on the interpretation of results, but also on the actual generation of results. It is concluded, as supported for example by Finnveden et al. (2007) and Ekvall et al. (2007), that no environmental system analysis study can provide undisputable, clear-cut solutions or completely objective information. This holds probably even more when (objective) statistical information is absent, as it typically is in MFA. In order to increase the confidence of decision makers in MFA results, it is important to communicate more than aggregate results on a material level, but to also provide relevant information on the informational basis of MFA systems. This can be facilitated by the procedures proposed in this thesis. The aggregated measures for information content can be used to communicate whether MFA results are reliable, and the uncertainty per flow can be used to indicate weaknesses in the systems (see Figure 30 and appendix 7). As to the comprehensive documentation in the DCM, the aggregated results can, if desired, easily be broken down to a level of detail making transparently visible that they depend on evaluations of individual attributes of data elements used in an MFA.

The procedure proposed for quantitative evaluation of information content and system structure (paragraph 2.3) may also be applied in other areas of the field of Industrial Ecology. In principle, it can be used for analysis of all systems that can be represented as networks, such as life cycle inventories or input-output models. In MFA, it complements existing methodologies by supporting system design, optimized use of available information and communication of MFA results. Issues of data quality and system structure, which have been qualitatively discussed in previous research, can now be gauged and quantitatively compared by means of the measures proposed in this thesis. Despite these possibilities, information content and system design in MFA are, in the presence of limited information, inescapably subjective to a certain degree. This is both a limitation of and an incentive to utilizing the procedure proposed in this thesis, as it facilitates working with limited, nonstatistical information in a systematic and transparent way. It is hoped that, in the future, the proposed measures will quantify increasing information content of MFAs over time and help to increase the acceptance of MFA results in scientific and institutional contexts.

References

Adriaans, P. 2013. Information. In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta: Stanford University.

Allenby, B. 2009. The industrial ecology of emerging technologies. *Journal of Industrial Ecology* 13(2): 168-183.

Arndt, C. 2004. *Information Measures: Information and Its Description in Science and Engineering*: Springer.

ASI. 2005. Stoffflussanalyse. In *Tell 1: Anwendung in der Abfallwirtschaft - Begriffe; Tell 2: Anwendung In der Abfallwirtschaft - Methodik; Teil 3: Stoffflussanalyse - Vorgangsweise bei der Bewertung*. Wien: Austrian Standards Institute.

Ayres, R. U. 1994. Industrial metabolism: theory and policy. In *Industrial Metabolism: Restructuring for Sustainable Development*, edited by R. U. Ayres, Simonis, U.K. . Tokyo: United Nations University Press.

Baccini, P. and P. H. Brunner. 1991. *Metabolism of the Anthroposphere*: Springer-Verlag.

Baccini, P. and H. P. Bader. 1996. *Regionaler Stoffhaushalt: Erfassung, Bewertung und Steuerung*: Spektrum-Akademischer Vlg.

Baccini, P. and P. H. Brunner. 2012. *Metabolism of the anthroposphere*. Cambridge: MIT Press.

Bailey, R., B. Bras, and J. K. Allen. 2004. Applying Ecological Input-Output Flow Analysis to Material Flows in Industrial Systems: Part II: Flow Metrics. *Journal of Industrial Ecology* 8(1-2): 69-91.

Baird, D. and R. E. Ulanowicz. 1989. The Seasonal Dynamics of The Chesapeake Bay Ecosystem. *Ecological Monographs* 59(4): 329-364.

Berger, J. O. and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76(2): 159-165.

Bettencourt, L. M. A. and C. Brelsford. 2015. Industrial Ecology: The View From Complex Systems. *Journal of Industrial Ecology* 19(2): 195-197.

Björklund, A. E. 2002. Survey of Approaches to Improve Reliability in LCA.". *The International Journal of Life Cycle Assessment* 7(2): 64-72.

Boltzmann, L. 1872. *Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen*. Vol. 66, *Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Klasse der Kaiserlichen Akademie der Wissenschaften*. Vienna.

Bonnin, M., C. Azzaro-Pantel, L. Pibouleau, S. Domenech, and J. Villeneuve. 2013. Development and validation of a dynamic material flow analysis model for French copper cycle. *Chemical Engineering Research and Design* 91(8): 1390-1402.

Brunner, P. H. and H. Rechberger. 2004. *Practical Handbook of Material Flow Analysis*. Boca Raton: Lewis Publishers.

Brunner, P. H. and H. Rechberger. 2016. *Handbook of Material Flow Analysis: For Environmental, Resource, and Waste Engineers, Second Edition*: CRC Press.

Buchner, H., D. Laner, H. Rechberger, and J. Fellner. 2014. In-depth analysis of aluminum flows in Austria as a basis to increase resource efficiency. *Resources, Conservation and Recycling* 93(0): 112-123.

Capurro, R. and B. Hjørland. 2003. The concept of information. *Annual Review of Information Science and Technology* 37(1): 343-411.

Cencic, O. 2016a. Nonlinear Data reconciliation in Material Flow Analysis with Software STAN. *Sustainable Environment Research*.

Cencic, O. 2016b. Treatment of Data Uncertainties in MFA. In *Handbook of Material Flow Analysis: For Environmental, Resource, and Waste Engineers, Second Edition*, edited by P. H. Brunner and H. Rechberger. Boca Raton: CRC Press.

Cencic, O. and R. Frühwirth. 2015. A general framework for data reconciliation—Part I: Linear constraints. *Computers & Chemical Engineering* 75: 196-208.

Chen, W.-Q. and T. E. Graedel. 2012. Anthropogenic Cycles of the Elements: A Critical Review. *Environmental Science & Technology* 46(16): 8574-8586.

Christian, R. R., D. Baird, J. Luczkovich, J. C. Johnson, U. M. Scharler, and R. E. Ulanowicz. 2005. Role of network analysis in comparative ecosystem ecology of estuaries. *Aquatic Food Webs* 3: 25e40.

Clark-Carter, D. 2014. Standard Error. In *Wiley StatsRef: Statistics Reference Online*: John Wiley & Sons, Ltd.

Clausius, R. 1867. *The Mechanical Theory of Heat: With Its Applications to the Steam-engine and to the Physical Properties of Bodies*: J. Van Voorst.

Clavreul, J., D. Guyonnet, D. Tonini, and T. Christensen. 2013. Stochastic and epistemic uncertainty propagation in LCA. *The International Journal of Life Cycle Assessment*: 1-11.

Côté, R. and J. Hall. 1995. Industrial parks as ecosystems. *Journal of Cleaner Production* 3(1–2): 41-46.

Danius, L., Burström, F. 2001. Regional material flow analysis and data uncertainties: Can the results be trusted? In *Sustainability in the Information Society. Part 2: Methods/Workshop Paper*, edited by L. M. G. Hilti, P.W. Marburg (D): Metropolis Verlag.

De Finetti, B. 1974. *Theory of probability: a critical introductory treatment*: John Wiley & Sons Australia, Limited.

Dijkema, G. P. J. and L. Basson. 2009. Complexity and Industrial Ecology. *Journal of Industrial Ecology* 13(2): 157-164.

Dijkema, G. P. J., M. Xu, S. Derrible, and R. Lifset. 2015. Complexity in Industrial Ecology: Models, Analysis, and Actions. *Journal of Industrial Ecology* 19(2): 189-194.

Dubois, D. and H. Prade. 2010. Formal Representations of Uncertainty. In *Decision-making Process: Concepts and Methods*: Wiley.

Dubois, D. and D. Guyonnet. 2011. Risk-informed decision-making in the presence of epistemic uncertainty. *International Journal of General Systems* 40(2): 145-167.

Džubur, N., O. Sunanta, and D. Laner. under revision. A fuzzy set-based approach for data reconciliation in material flow modeling *Applied Mathematical Modelling*.

Egle, L., O. Zoboli, S. Thaler, H. Rechberger, and M. Zessner. 2014. The Austrian P budget as a basis for resource optimization. *Resources, Conservation and Recycling* 83(0): 152-162.

Ekvall, T., G. Assefa, A. Björklund, O. Eriksson, and G. Finnveden. 2007. What life-cycle assessment does and does not do in assessments of waste management. *Waste Management* 27(8): 989-996.

Finnveden, G., A. Björklund, Å. Moberg, T. Ekvall, and Å. Moberg. 2007. Environmental and economic assessment methods for waste management decision-support: possibilities and limitations. *Waste Management & Research* 25(3): 263-269.

Fischer-Kowalski, M. 1998. Society's Metabolism. *Journal of Industrial Ecology* 2(1): 61-78.

Fisher, R. A. 1925. Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 22(05): 700-725.

Floridi, L. 2013. *The Philosophy of Information*: Oxford University Press.

Frosch, R. A. and N. E. Gallopoulos. 1989. Strategies for manufacturing. *Scientific American* 261(3): 144-152.

Funtowicz, S. O. and J. R. Ravetz. 1993. Science for the post-normal age. *Futures* 25(7): 739-755.

Goerner, S. J., B. Lietaer, and R. E. Ulanowicz. 2009. Quantifying economic sustainability: Implications for free-enterprise theory, policy and practice. *Ecological Economics* 69(1): 76-81.

Gottschalk, F., R. W. Scholz, and B. Nowack. 2010. Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles. *Environmental Modelling & Software* 25(3): 320-332.

Graedel, T. E. 1996. On the concept of Industrial Ecology. *Annual Review of Energy and the Environment* 21(1): 69-98.

Graedel, T. E., D. van Beers, M. Bertram, K. Fuse, R. B. Gordon, A. Gritsinin, A. Kapur, R. J. Klee, R. J. Lifset, L. Memon, H. Rechberger, S. Spatari, and D. Vexler. 2004. Multilevel Cycle of Anthropogenic Copper. *Environmental Science & Technology* 38(4): 1242-1252.

Habib, K., P. K. Schibye, A. P. Vestbo, O. Dall, and H. Wenzel. 2014. Material Flow Analysis of NdFeB Magnets for Denmark: A Comprehensive Waste Flow Sampling and Analysis Approach. *Environ Sci Technol* 48(20): 12229-12237.

Hedbrant, J. and L. Sörme. 2001. Data Vagueness and Uncertainties in Urban Heavy-Metal Data Collection. *Water, Air and Soil Pollution: Focus* 1(3-4): 43-53.

Heijungs, R. 2015. Topological network theory and its application to LCA and IOA and related industrial ecology tools: principles and promise. *J Environ Account Manag* 3(2): 151-167.

Heijungs, R. and M. A. J. Huijbregts. 2004. A review of approaches to treat uncertainty in LCA. In *Proceedings of the 2nd Biennial Meeting of iEMSs, Complexity and integrated resources management, 14-17 June 2004*. Osnabrück, Germany: Orlando, Fla. : Elsevier.

Hintikka, J. 1973. *Logic, Language-games and Information: Kantian Themes in the Philosophy of Logic*: Clarendon Press.

Hofko, B., M. Dimitrov, O. Schwab, F. Weiss, H. Rechberger, and H. Grothe. 2016. Technological and environmental performance of temperature-reduced mastic asphalt mixtures. *Road Materials and Pavement Design*: 1-16.

Huang, D.-B., H.-P. Bader, R. Scheidegger, R. Schertenleib, and W. Gujer. 2007. Confronting limitations: New solutions required for urban water management in Kunming City. *Journal of Environmental Management* 84(1): 49-61.

Kay, J. J. 2002. On complexity theory, exergy and industrial ecology. In *Construction ecology - Nature as the basis for green buildings*, edited by C. J. Kibert, et al. New York: Spon Press.

Kharrazi, A., E. Rovenskaya, B. D. Fath, M. Yarime, and S. Kraines. 2013. Quantifying the sustainability of economic resource networks: An ecological information-based approach. *Ecological Economics* 90: 177-186.

Klee, R. J. and T. E. Graedel. 2004. Elemental Cycles: A Status Report on Human or Natural Dominance. *Annual Review of Environment and Resources* 29(1): 69-107.

Klinglmair, M., O. Zoboli, D. Laner, H. Rechberger, T. F. Astrup, and C. Scheutz. 2016. The effect of data structure and model choices on MFA results: A comparison of phosphorus balances for Denmark and Austria. *Resources, Conservation and Recycling* 109: 166-175.

Kolmogorov, A. N. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2(1-4): 157-168.

Kopec, G. M., J. M. Allwood, J. M. Cullen, and D. Ralph. 2015. A General Nonlinear Least Squares Data Reconciliation and Estimation Method for Material Flow Analysis. *Journal of Industrial Ecology*: n/a-n/a.

Korhonen, J. 2001. Four ecosystem principles for an industrial ecosystem. *Journal of Cleaner Production* 9(3): 253-259.

Laner, D., H. Rechberger, and T. Astrup. 2014. Systematic Evaluation of Uncertainty in Material Flow Analysis. *Journal of Industrial Ecology* 18(6).

Laner, D., H. Rechberger, and T. Astrup. 2015a. Applying fuzzy and probabilistic uncertainty concepts to the material flow analysis of palladium in Austria. *Journal of Industrial Ecology* 19(6): 1055-1069.

Laner, D., J. Feketitsch, H. Rechberger, and J. Fellner. 2015b. A novel approach to characterize data uncertainty in MFA and its applications to plastic flows in Austria. *Journal of Industrial Ecology*.

Layton, A., B. Bras, and M. Weissburg. 2016. Industrial Ecosystems and Food Webs: An Expansion and Update of Existing Data for Eco-Industrial Parks and Understanding the Ecological Food Webs They Wish to Mimic. *Journal of Industrial Ecology* 20(1): 85-98.

Lazarevic, D., N. Buclet, and N. Brandt. 2012. The application of life cycle thinking in the context of European waste policy. *Journal of Cleaner Production* 29–30: 199-207.

Lenzen, M., R. Wood, and T. Wiedmann. 2010. Uncertainty Analysis for Multi-Region Input-Output Models - A Case Study of the UK's Carbon Footprint. *Economic Systems Research* 22(1): 43-63.

Liu, G. and D. B. Müller. 2013. Mapping the Global Journey of Anthropogenic Aluminum: A Trade-Linked Multilevel Material Flow Analysis. *Environmental Science & Technology* 47(20): 11873-11881.

Lloyd, S. M. and R. Ries. 2007. Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment: A Survey of Quantitative Approaches. *Journal of Industrial Ecology* 11(1): 161-179.

Madnick, S. and H. Zhu. 2006. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering* 59(2): 460-475.

Mao, J. S., J. Dong, and T. E. Graedel. 2008. The multilevel cycle of anthropogenic lead: I. Methodology. *Resources, Conservation and Recycling* 52(8–9): 1058-1064.

McMullin, E. 1968. What do Physical Models Tell us? In *Studies in Logic and the Foundations of Mathematics*, edited by B. V. Rootselaar and J. F. Staal: Elsevier.

Meerow, S. and J. P. Newell. 2015. Resilience and Complexity: A Bibliometric Review and Prospects for Industrial Ecology. *Journal of Industrial Ecology* 19(2): 236-251.

Morgan, M. G., M. Henrion, and M. Small. 1992. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*: Cambridge University Press.

Nakajima, K., H. Ohno, Y. Kondo, K. Matsubae, O. Takeda, T. Miki, S. Nakamura, and T. Nagasaka. 2013. Simultaneous Material Flow Analysis of Nickel, Chromium, and Molybdenum Used in Alloy Steel by Means of Input–Output Analysis. *Environmental Science & Technology* 47(9): 4653-4660.

Navarrete-Gutiérrez, T., B. Rugani, Y. Pigné, A. Marvuglia, and E. Benetto. 2015. On the Complexity of Life Cycle Inventory Networks: Role of Life Cycle Processes with Network Analysis. *Journal of Industrial Ecology*: n/a-n/a.

Nuss, P., W.-Q. Chen, H. Ohno, and T. E. Graedel. 2016. Structural Investigation of Aluminum in the U.S. Economy using Network Analysis. *Environmental Science & Technology* 50(7): 4091-4101.

Odum, H. T. 1994. *Ecological and General Systems: An Introduction to Systems Ecology*: University Press of Colorado.

Ott, C. and H. Rechberger. 2012. The European phosphorus balance. *Resources, Conservation and Recycling* 60: 159-172.

Rechberger, H. 2015. Die Geheimhaltung von Daten als Hindernis für die sog "Nationale Rohstoffbuchhaltung". In *Jahrbuch Abfallwirtschaftsrecht 2015*, edited by C. Piska and B. Lindner: Neuer Wissenschaftlicher Verlag.

Reck, B. K., M. Chambon, S. Hashimoto, and T. E. Graedel. 2010. Global Stainless Steel Cycle Exemplifies China's Rise to Metal Dominance. *Environmental Science & Technology* 44(10): 3940-3946.

Refsgaard, J. C., J. P. van der Sluijs, A. L. Højberg, and P. A. Vanrolleghem. 2007. Uncertainty in the environmental modelling process – A framework and guidance. *Environmental Modelling & Software* 22(11): 1543-1556.

Rutledge, R. W., B. L. Basore, and R. J. Mulholland. 1976. Ecological stability: An information theory viewpoint. *Journal of Theoretical Biology* 57(2): 355-371.

Schiller, F., A. S. Penn, and L. Basson. 2014. Analyzing networks in industrial ecology – a review of Social-Material Network Analyses. *Journal of Cleaner Production* 76: 1-11.

Schwab, O. and H. Rechberger. 2014. *Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen - Teil 1: Analyse der Daten und Quellen (Projekt EDNA 1)*. Report for the Austrian Federal Ministry of Science, Research and Economy, Technische Universität Wien, Vienna.

Schwab, O. and H. Rechberger. 2015. *Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen - Teil 3: Entwicklung von SFA-spezifischen Datenqualitätsindikatoren (Projekt EDNA3)*. Report for the Austrian Federal Ministry of Science, Research and Economy, Technische Universität Wien, Vienna.

Schwab, O. and H. Rechberger. under revision. Information content, Complexity and Uncertainty in Material Flow Analysis. *Journal of Industrial Ecology*.

Schwab, O., O. Zoboli, and H. Rechberger. 2016a. A Data Characterization Framework for Material Flow Analysis. *Journal of Industrial Ecology*.

Schwab, O., D. Laner, and H. Rechberger. 2016b. Quantitative evaluation of data quality in regional Material Flow Analysis. *Journal of Industrial Ecology*.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27: 379-423.

Shannon, C. E. and W. Weaver. 1963. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Suh, S. 2005. Theory of materials and energy flow analysis in ecology and economics. *Ecological Modelling* 189(3–4): 251-269.

Trinkel, V., N. Kieberger, T. Bürgler, H. Rechberger, and J. Fellner. 2015. Influence of waste plastic utilisation in blast furnace on heavy metal emissions. *Journal of Cleaner Production* 94: 312-320.

Ulanowicz, R. E. 1980. An hypothesis on the development of natural communities. *Journal of Theoretical Biology* 85(2): 223-245.

Ulanowicz, R. E. 1997. *Ecology, the Ascendent Perspective*: Columbia University Press.

Ulanowicz, R. E., S. J. Goerner, B. Lietaer, and R. Gomez. 2009. Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity* 6(1): 27-36.

Vadenbo, C., G. Guillén-Gosálbez, D. Saner, and S. Hellweg. 2014. Multi-objective optimization of waste and resource management in industrial networks – Part II: Model application to the treatment of sewage sludge. *Resources, Conservation and Recycling* 89: 41-51.

van Eygen, E., J. Feketitsch, D. Laner, H. Rechberger, and J. Fellner. 2016. Comprehensive analysis and quantification of national plastic flows: the case of Austria. *Resources, Conservation and Recycling*.

von Neumann, J. 1955. *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer.

Walker, W. E., P. Harremoës, J. Rotmans, J. P. van der Sluijs, M. B. A. van Asselt, P. Janssen, and M. P. Krayer von Krauss. 2003. Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment* 4(1): 5-17.

Weidema, B. P. and M. S. Wesnæs. 1996. Data quality management for life cycle inventories—an example of using data quality indicators. *Journal of Cleaner Production* 4(3–4): 167-174.

Wiedmann, T., H. C. Wilting, M. Lenzen, S. Lutter, and V. Palm. 2011. Quo Vadis MRIO? Methodological, data and institutional requirements for multi-region input–output analysis. *Ecological Economics* 70(11): 1937-1945.

Wolman, A. 1965. The metabolism of cities. *Scientific American* 213(3): 179-190.

Wood, R. and M. Lenzen. 2009. Aggregate Measures of Complex Economic Structure and Evolution. *Journal of Industrial Ecology* 13(2): 264-283.

Wulff, F., J. G. Field, and K. H. Mann. 1989. *Network analysis in Marine Ecology: Methods and Applications.*: Springer.

Zoboli, O., M. Zessner, and H. Rechberger. 2015. Added Value of Time Series in MFA: The Austrian Phosphorus Budget from 1990 to 2011. *Journal of Industrial Ecology*.

Zoboli, O., M. Zessner, and H. Rechberger. 2016. Supporting phosphorus management in Austria: Potential, priorities and limitations. *Science of the total Environment* 565: 313-323.

List of figures

List of tables

Glossary

Complexity (system)	A characteristic of a material flow system. A complex system has, as opposed to a trivial system, a non-trivial network topology.
Data attributes	Data-associated annotations concerning statistical properties, meaning, origination and application of the data.
Database analysis	A three-step procedure for analysis of MFA databases, consisting of data inventory, evaluation of data elements and analysis of data attributes.
Data characterization matrix	A template for systematic characterization of MFA input data and basis for data quality evaluation.
Data element	Representation of an entity by a numeric value (data point) or by more than one numeric values (interval, data set).
Data quality	Designates if data is "fit for purpose". Data quality may be expressed by statistical measures or evaluated according to the ability of data to meet certain criteria.
Data uncertainty	A state of being uncertain about a phenomenon or event as a consequence of limited or missing information.
Entity	A real-world phenomenon or real-world object.
Information	Data plus meaning.
Information background	Origination and forming process of a piece of information.
Information content	Difference between the uncertainty of a quantitative MFA system and the uncertainty of a qualitative MFA system.
Information defects	Information shortcomings that reduce the degree of belief in a given piece of information to be true in a particular context.
Information element	A piece of information that represents an entity. An information element can consist of one or several data elements.
Information level	Information in MFA can be described on four levels (data element, information, information background and MFA information).
MFA information	Information with background that is put into context of an MFA study.
System elements	The components of material flow systems ("flows", "processes", "stocks", and "materials").
Triviality (system)	A characteristic of a material flow system. A trivial system has, as opposed to a complex system, a trivial network topology.

List of abbreviations

A	Absolute; a measurement scale of data attributes (“numbers”)
B	Binary; a measurement scale of data attributes (“yes or no”)
C	Structural complexity of a system
DCM	Data characterization matrix
Fi	Flow *i*
ID	Information defect
k	Positive constant
N	Nominal; a measurement scale of data attributes (“in words”)
O	Ordinal; a measurement scale of data attributes (“between 0 and 1”)
P	Process
T	Structural triviality of a system
S	“Informational” size of a system
U	Uncertainty of a system
X	Quantity of a flow
y	Source process of a flow
z	Target process of a flow
Subscripts	
ap	A priori
b	Balanced
Fi	Property of flow *Fi*
S, R, P, C	Information defects “semantic”, “representativeness”, “provenance” and “context”. A set of the four information defects *S*, *R*, *P* and *C* is designated as ID_i.
tot	Total Information defect of an information element (combination of ID_S, ID_R, ID_P and ID_C)
w	Weighted

Appendences

Appendix 1: Data characterization matrices including computation of information defects, information content and system structure

The following Excel spreadsheets are provided online in the TU Catalog Plus (www.ub.tuwien.ac.at) and can be accessed via the library entry of this dissertation.

A1.1: Phosphorus
A1.2: Palladium
A1.3: Plastics
A1.4: Aluminum

Appendix 2: Attributes of the data characterization matrix (DCM)

Information level		Attribute group	Attribute name	Number	Scale	Description
1	data element	stat. characteristics	data element form	101	N	What form does the data element have?
			location parameter	102	N	Which location parameter is provided and introduced in the MFA?
			value (numeric)	103	A	What is the numeric value of this location parameter?
			n	104	A	What is the total number of samples?
			min	105	A	What is the min value?
			max	106	A	What is the max value?
			distribution (form)	107	N	What form does the probability distribution have?
			distribution (paramet.)	108	A	Which numeric values do the parameters of the distribution have?
			dispersion (measure)	109	N	Which dispersion measure is provided?
			dispersion (numeric)	110	A	Which numeric value does the dispersion measure have?
2	information	semantics	description of meaning	201	N	What is the meaning of the entity, described in natural language?
			semantic precision	202	O	Is the meaning of the entity precisely know and linguistically precise, or ambiguous or imprecise?
		scale	entity category	203	N	Which category does the entity belong to?
			entity class	204	N	Which class does the entity belong to?
			unit	205	N	What is the unit of the entity?
			sphere	206	N	Is the entity part of the "anthroposphere" or part of the "natural environment"?
			property type	207	N	Is the data element extensive (dependent on the system size) or intensive (independent from the system size)?
			mathematical form	208	N	Is the entity discrete or continuous?
			minimum (potential)	209	A	What is the lowest possible value the data element could potentially have?
			maximum (potential)	210	A	What is the highest possible value the data element could potentially have?
		complexity	variety	211	O	Does the entity refer to a vast number of possible real-world objects or to one particular object?
			disparity	212	O	How different are the real-world objects an entity refers to (e.g. "mobile phones" refers to a large number of different objects)?
3	background information	availability	existence	301	B	Does the data element exist or not?
			accessibility	302	B	Is the data element accessible?
			access restriction	303	N	If not accessible, why not?
		communication	communication type	304	N	How is the data element communicated?
			access type	305	N	How can the data element be accessed?
			frequency	306	N	How frequently is the data updated?
		producer	producer category	307	N	In which sector was the data element produced?
			producer type	308	N	Which institution did produce the data element?
			reference	309	N	What is the specific reference?
		origination	origination category	310	N	Was the data element empirically collected or derived (from data, assumptions, speculation)?
			origination type	311	N	How was the data collected?
			origination type quality	312	O	How is the quality of the origination type (precision of an empirical method, quality of a model, expertise of an estimator)?
4	MFA information	application in MFA	utilization type	401	N	Is the regarded data element applied for the description of a flow, a process, a stock or others?
			autonomy	402	O	Can the data element be applied in the study directly, or must it be combined with more data before?
			layer	403	N	Was this data element collected for quantification on the goods or on the substance layer?
			type of good	404	N	What kind of good does it refer to?
		system relation	primary determination	405	N	Which sphere does primarily determine the value?
			temporal variability	406	O	How much does the value vary over time?
			trend	407	N	Is there a systematic temporal relation?
			spatial variability	408	O	How variable is the data over space?
			further relation	409	N	Is there, in addition to time and space, another relevant relation that could influence the adequacy of the data element?
			variability by further relation	410	O	How variable is the data element by this further relation?
		system adequacy	temporal divergence	411	A	How well is the data element within the temporal system boundary?
			spatial divergence	412	O	How well is the data element within the spatial system boundary?
			further divergence	413	O	Is there a further divergence, (e.g. does the data element describe a similar, but different, process)?
			adaptation (type)	414	N	How was the data element adapted (e.g. scaled)?
			adaptation (quality)	415	O	How well does a data element after adjustment fit the system?
			missing values			NA

Appendix 3: Code for the data characterization matrix (DCM)

BINARY ATTRIBUTES (yes/no) information level	attribute group	attribute	no	attribute values	
background information	availability	existence	301	The wanted data exist.	The wanted data do not exist.
background information	availability	accessibility	302	The wanted data are available.	The wanted data are not available.

ORDINAL ATTRIBUTES (ranked between 0 and 1) *information level*	*attribute group*	*attribute*	*no.*	*attribute values* *0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*1*
information	*semantics*	*semantic precision*	*202*	*The entity is described in natural language in an unambiguous and precise way.*										*The description of the entity in natural language is strongly ambiguous and imprecise.*
	Complexity	*variety*	*211*	*The data relate to a specific real-world entity, e.g. " VW Golf 1.6, by 2006", or "iPhone 5"*										*The data relate to a vast group of entities of a similar class, e.g. "WEEE"*
		disparity	*212*	*The values that describe objects or phenomena are homogenous. Example: Al-content of beverage cans*										*The values to describe objects or phenomena are very widely spread or not manageable (heterogeneous). Example: Cu-content of mobile phones.*
information background	*origination*	*origination type quality*	*312*	*The applied data collection method was conducted with good quality (precise measurement, estimator with a high level of expertise, good access to relevant information of an agent).*										*The applied data collection method was conducted with poor quality (imprecise measurement, estimator with poor level of expertise, scarce access to relevant information of an agent).*
MFA information	*application in MFA*	*autonomy*	*402*	*The data are applied autonomously in the study.*										*The data are applied in combination with a vast number of additional data elements.*
	system relation	*temporal variability*	*406*	*The data do not vary (are constant) over time.*										*The data vary strongly over time.*
		spatial variability	*408*	*The data do not vary (are constant) over space.*										*The data vary strongly over space.*
		variability by further relation	*410*	*The data do not vary by a further relation.*	*Little variability by further relation, for example for standardized technical processes, general socio-cultural processes or globalized markets.*									*High variability by further relation, for example for highly specialized processes (e.g. high tech), specific socio-cultural processes (e.g. consumer behavior), or markets (niche markets).*
	system adequacy	*spatial divergence*	*412*	*The data fit the spatial system boundary.*	*The data are on a different, but very similar spatial system (e.g. similar geography, nation with similar development status).*									*The data are on a very different spatial system, or the location the data relate to is not specified.*
		further divergence	*413*	*The data fit the system, i.e. are not divergent by any further relation.*	*The data are hardly divergent from the studied system (similar technical process, similar socio-cultural process, similar market).*									*The data are strongly divergent from the studied system (e.g. very different technical process, socio-cultural or market process).*
		adaptation (quality)	*415*	*The data are adequate to the studied system and no adaption is necessary, or data after adjustment are considered accurate to the boundary of the studied system.*										*Adaptations to fit the data to the studied system are very rough or speculative.*

ABSOLUTE ATTRIBUTES (numbers)				
information level	attribute group	attribute	no.	attribute values
data element	statistical characteristics	value (numeric)	103	numeric
		n	104	numeric
		min	105	numeric
		max	106	numeric
		distribution (param.)	108	numeric
		dispersion (numeric)	110	numeric
information	Scale	minimum (pot.)	209	numeric
		maximum (pot.)	210	numeric
MFA information	system adequacy	temporal divergence	411	numeric

NOMINAL ATTRIBUTES (text)				
information level	attribute group	attribute	no	attribute specifications
		data element form	101	point, set, interval
		location parameter	102	mean, mode, median, unspecified
		distribution	107	normal, lognormal, gamma, betageneral, weibull, dirac, uniform, triang, other
		dispersion (measure)	109	standard deviation, variance, standard error, uncert. interval (±), uncert. interval (*/)
information	semantics	descr. of meaning	201	Description in natural language.
	scale	entity category	203	space, time, rate (x/y; time-related), quota (x/y), quantity
		entity class	204	number, mass, mass/area, mass/time, mass/mass, mass/volume, mass/item, concentration (mass.-%), area, volume, number/time, volume/time, mass/cross section, mass/item
		unit	205	dimensionless, SI-unit
		sphere	206	anthroposphere, natural environment
		property type	207	intensive, extensive
		mathematical form	208	continuous, discrete
information background	availability	access restriction	303	secrecy, costs, none
	communicat.	communication type	304	public, on request, conditioned, none
		access type	305	online, report, book, journal publication, proceedings, legislative document, personal communication, none
		frequency	306	one-off source, annually, biannually, five-annually, ten-annually, permanently, irregular
	producer	producer category	307	authority, science, economy, civil society
		producer type	308	statistical office, ministry, administration, university, non-university, company, industrial association, interest group, NGO, association
		reference	309	name of the institution/ Reference
	origination	origination category	310	empirical, derived (mainly from data), derived (mainly from assumptions), derived (from speculation)
		origination type	311	counting, measurement (lab), measurement (in-situ), industrial monitoring, accounting, reporting, expert estimation, speculation, assumption, model, calculation, planning documents, survey
MFA information	utilization in MFA	utilization type	401	flow, flux, process, material, stock, stock change rate, transfer coefficient, precursor
		layer	403	good, substance, none
		type of good	404	raw material, industrial good, capital good, consumer good, durable consumer good, infrastructure, living organism, emission, waste, none
	system relation	primary determination	405	technology, market, sociosphere, political decisions, biosphere, geosphere, scientific rationale
		trend	407	increasing, decreasing, fluctuating, constant, none
		further relation	409	technology, market, interpretation, reference unit, none
	system adequacy	adaptation (type)	414	none, scaling (temporal), scaling (spatial), scaling (to system size), conversion

Appendix 4: Scheme for translation of data attributes to mathematically computable scales

Table A4.1: Attributes that need to be translated for application in the information defect functions

Attr. No	Name	Translation activity	Original scale	Translated to
a308	Producer type	Ranking the information producers from 0 (reliable) to 1 (unreliable). The producer type specification "science" is sub-divided by merging a308 with a305 (access type).	Nominal	Ordinal
a311	Origination type	Ranking the origination types from 0 (good origination method) to 1 (poor method).	Nominal	Ordinal
a411	Temporal divergence	Categorizing years of divergence on a scale from 0 (little divergence) to 1 (high divergence).	Absolute	Ordinal

Table A4.2: Default translation of attributes applied in the case studies

a308 (producer type)		a311 (origination type)		a411 (temporal divergence)	
Nominal	Ordinal	Nominal	Ordinal	Absolute	Ordinal
Statistical office	0	Counting	0	0	0
Ministry	0.1	Measurement (lab)	0.1	1	0.1
Administration	0.1	Measurem. (in-situ)	0.1	2	0.2
Book	0.1	Industrial monitoring	0.1	3	0.3
Journal publication	0.2	Accounting	0.1	4	0.4
Proceedings	0.3	Reporting	0.1	5	0.5
Report	0.3	Calculation	0.2	6	0.6
Personal commun.	0.4	Model	0.3	7	0.7
Industrial association	0.4	Survey	0.3	8	0.8
Company	0.5	Expert estimation	0.4	9	0.9
Association	0.8	Planning documents	0.6	≥10	1
NGO	0.9	Legislative document	0.8		
Interest group	0.9	Speculation	1		
None/NA	1				

Appendix 5: Surface plots of the information defect functions

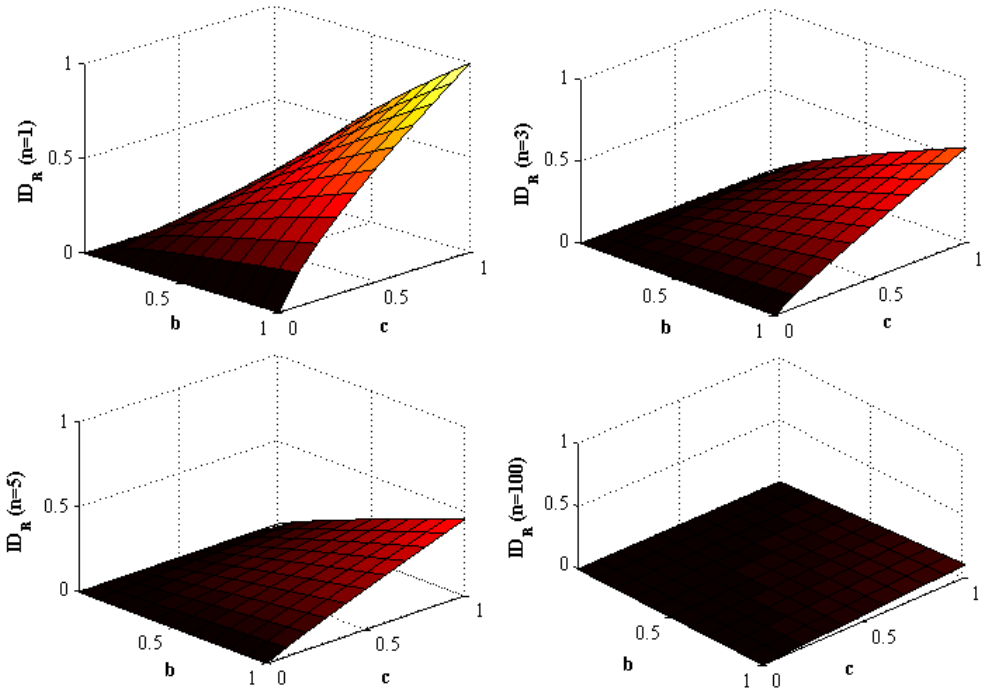


Figure A5.1: Representativeness information defect $ID_R=f(b,c,n)$.

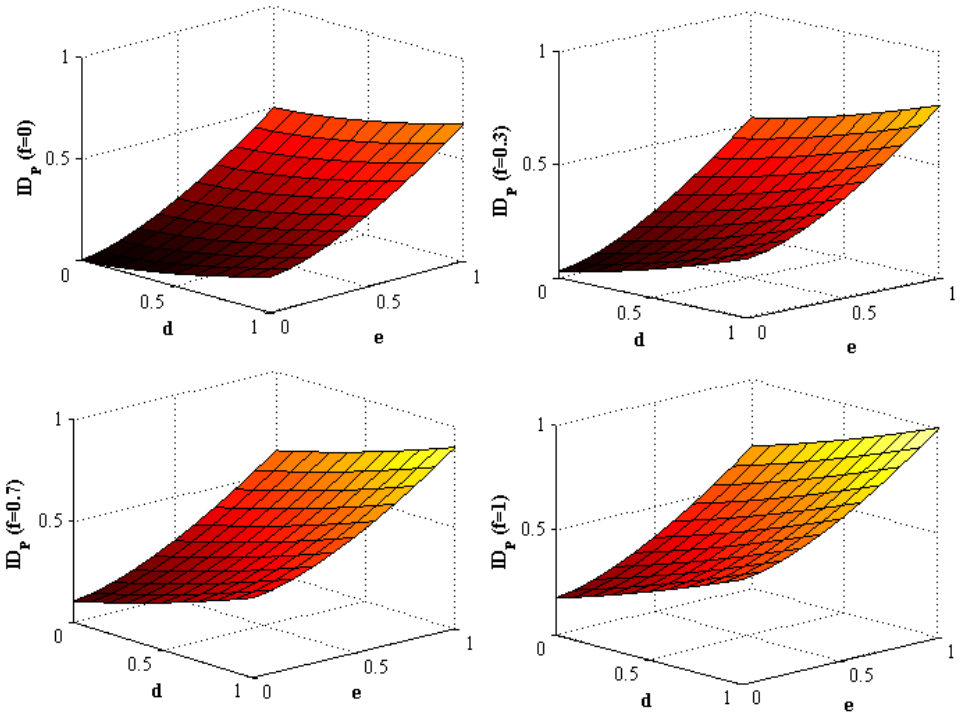


Figure A5.2: Provenance information defect $ID_P=f(d,e,f)$.

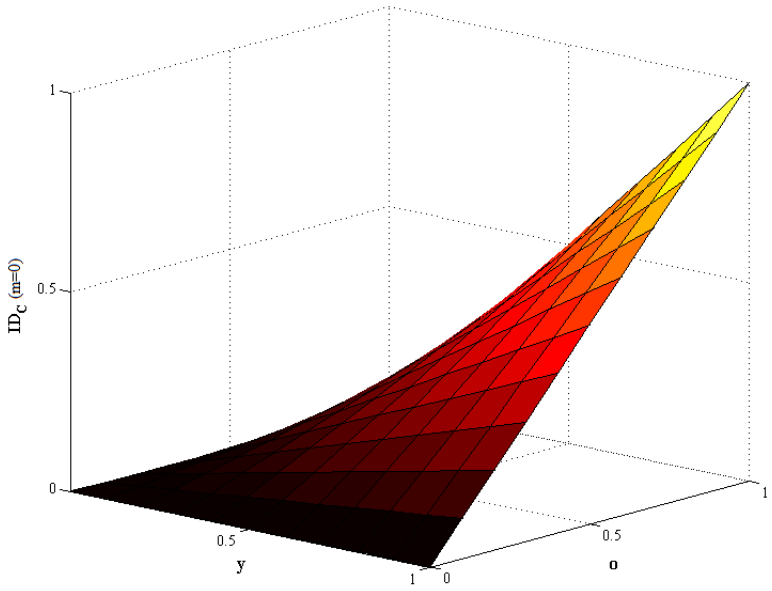


Figure A5.3: Context information defect $ID_C=f(y,o,m)$. m is a binary attribute (0,1) and only the case m=0 is plotted. If m=1, it is $ID_C=y$.

Appendix 6: Graphical comparison of two ID_F normalization functions

Two functions for normalization of ID_{tot} to the measurement scale (0-1) are compared, where z is the number of information elements per flow (z=1, 2, 3…) which is most often 2 and typically never higher than 4.

$$SQRT: ID_F = \frac{\sqrt{\sum_{i=1}^{z} ID_{tot,i}{}^2}}{\sqrt{z}}$$

and

$$\text{LOGISTIC: } ID_F = \frac{1.5}{(1+2e^{-3\sqrt{\sum_{i=1}^{z} ID_{tot,i}{}^2}})} - 0.5$$

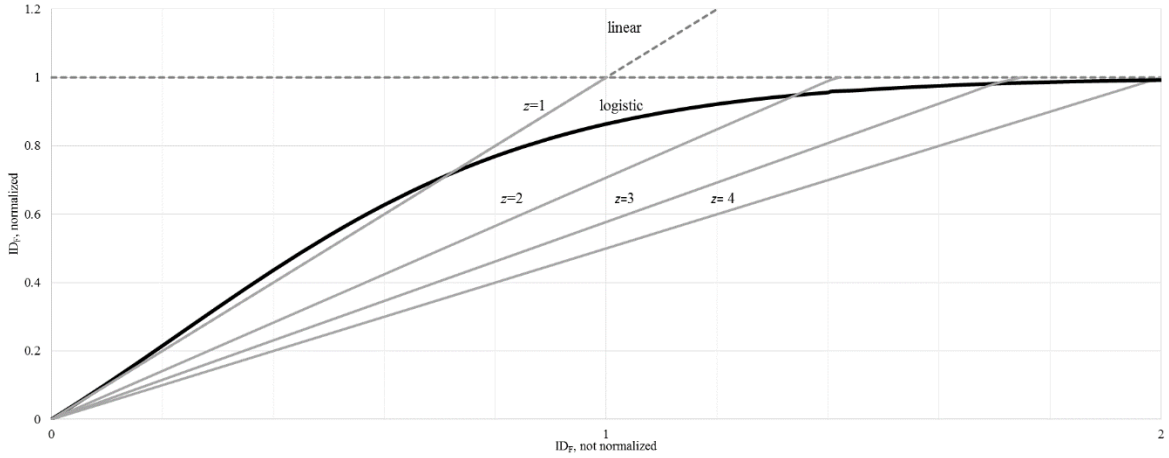


Figure A6.1: Comparison of ID_F normalization by a logistic function (black) and by the square root of the number of information elements z (grey).

It depends on the interpretation of "information defect" if LOGISTIC or SQRT is chosen for normalization of ID_F. If information defects are regarded as "intensive quantities" (such as concentrations, which are intrinsic properties of a system and not dependent on the system size), SQRT should be chosen. SQRT averages the information defects of multiple ID_{tot} per flow. It does not assume ID_F to increase with increasing number of information elements. If information defects are regarded as degrees of believe in information to be true, LOGISTIC should be chosen. It accumulates the information defects of multiple ID_{tot} per flow. It is differentiating for low information defects (see the quasi-linear first part of the black line in Figure A6.1). The lower the degree of belief in information to be true (i.e. the higher the information defect), the less differentiating is LOGISTIC. This reflects the assumption "the more vague information is, the lower are possibilities of agents to distinguish between two pieces of information". The difference between SQRT and LOGISTIC is specified in two examples.

(1) SQRT averages, LOGISTIC accumulates

For a flow with two information elements ($ID_{tot,1}$ = 0.2 and $ID_{tot,2}$ = 0.8), LOGISTIC delivers ID_F = 0.69 and SQRT delivers ID_F = 0.48. SQRT produces an averaged value and enables consistent linear scaling of ID_F throughout the measurement scale. The result of LOGISTIC is never lower than the highest defect of the information elements used for its calculation until the point of intersection between LOGISTIC and SQRT with z=1. In the above proposed parametrization of the logistic function, this is at $\sqrt{\sum_{i=1}^{z} ID_{tot,i}{}^2} = 0.7$. This means that for defects above the intersection that diverge towards one, LOGISTIC is decreasingly differentiating (see Figure A6.1).

(2) SQRT is not sensitive to the number of information elements, LOGISTIC is

In the Pd case study, SQRT delivers lower results than LOGISTIC for all z>1 (see Figure A6.2). This may affect the ranking of information defects (see for example flow F19, where z=3). SQRT is not sensitive to z and ID_F increases only with increasing ID_{tot} of the information elements. LOGISTIC is sensitive to the number of information elements, and ID_F increases with z. For a flow with two information elements, both with a defect ID_{tot}=0.3, SQRT delivers ID_F=0.3 and LOGISTIC delivers ID_F=*0.46*. For three information elements with identical ID_{tot}=0.3, SQRT delivers ID_F=0.3, LOGISTIC delivers ID_F=0.56. LOGISTIC means "the more vague information elements z per flow, the higher the information defect ID_F", SQRT means "the number of information elements does not influence the information defect".

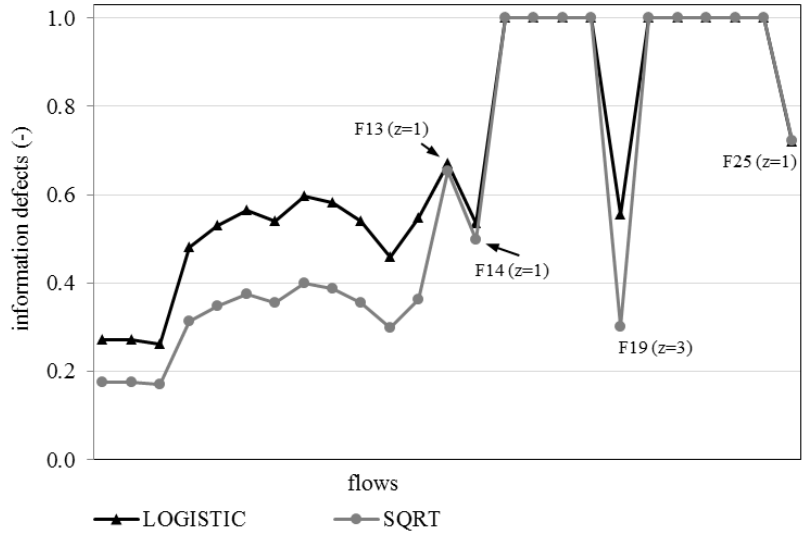


Figure A6.2: Application of two ID_F normalization options (SQRT and LOGISTIC) to the Pd case study.

Appendix 7: Additional flowcharts visualizing the uncertainty and complexity of the aluminum and plastics systems

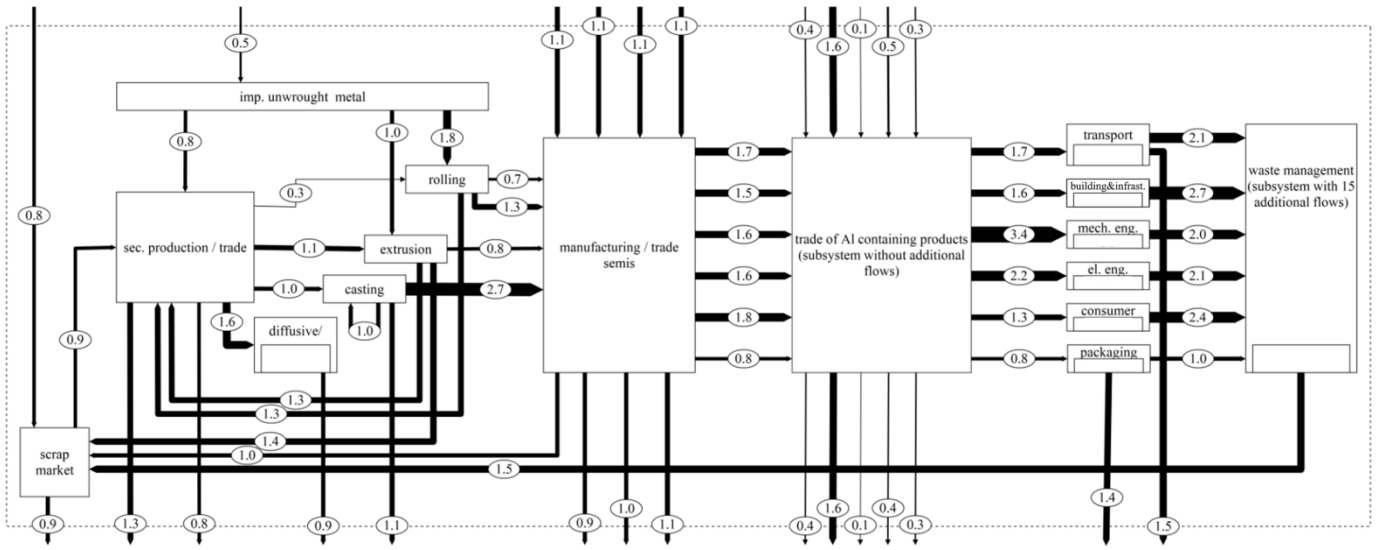


Figure A7.1: Uncertainty U_b of the flows in the aluminum MFA displayed as a flowchart. Flow widths are proportional to the uncertainty per balanced flow. For the total system, it is U_b=99.

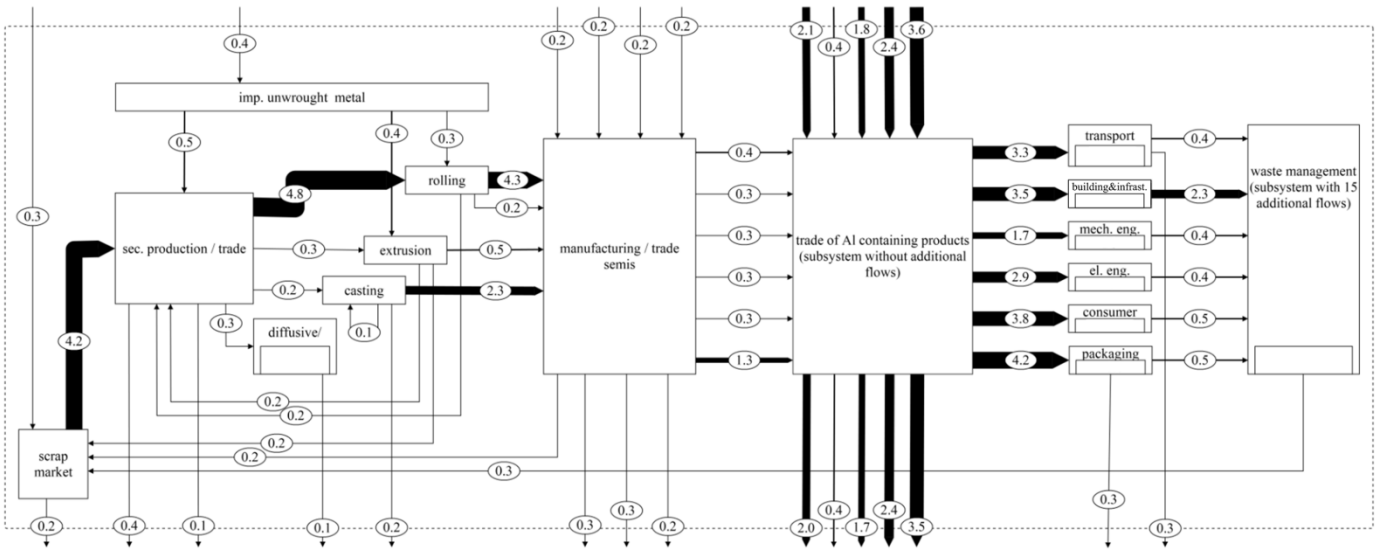


Figure A7.2: Difference between U_{ap} and U_b of the flows in the aluminum MFA displayed as a flowchart, indicating the degree of data reconciliation per flow.

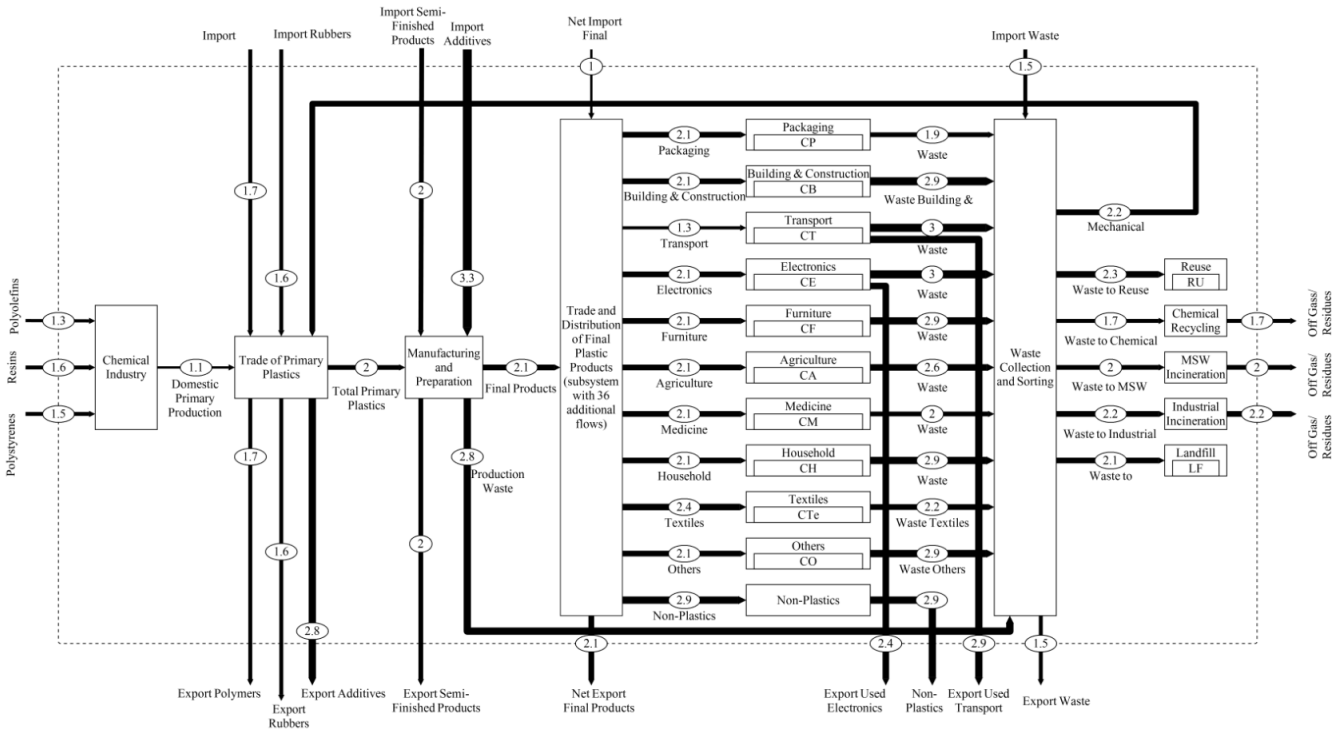


Figure A7.3: Uncertainty U_b of the flows in the plastics MFA displayed as a flowchart. Flow widths are proportional to the uncertainty per balanced flow. For the total system, it is U_b=202.

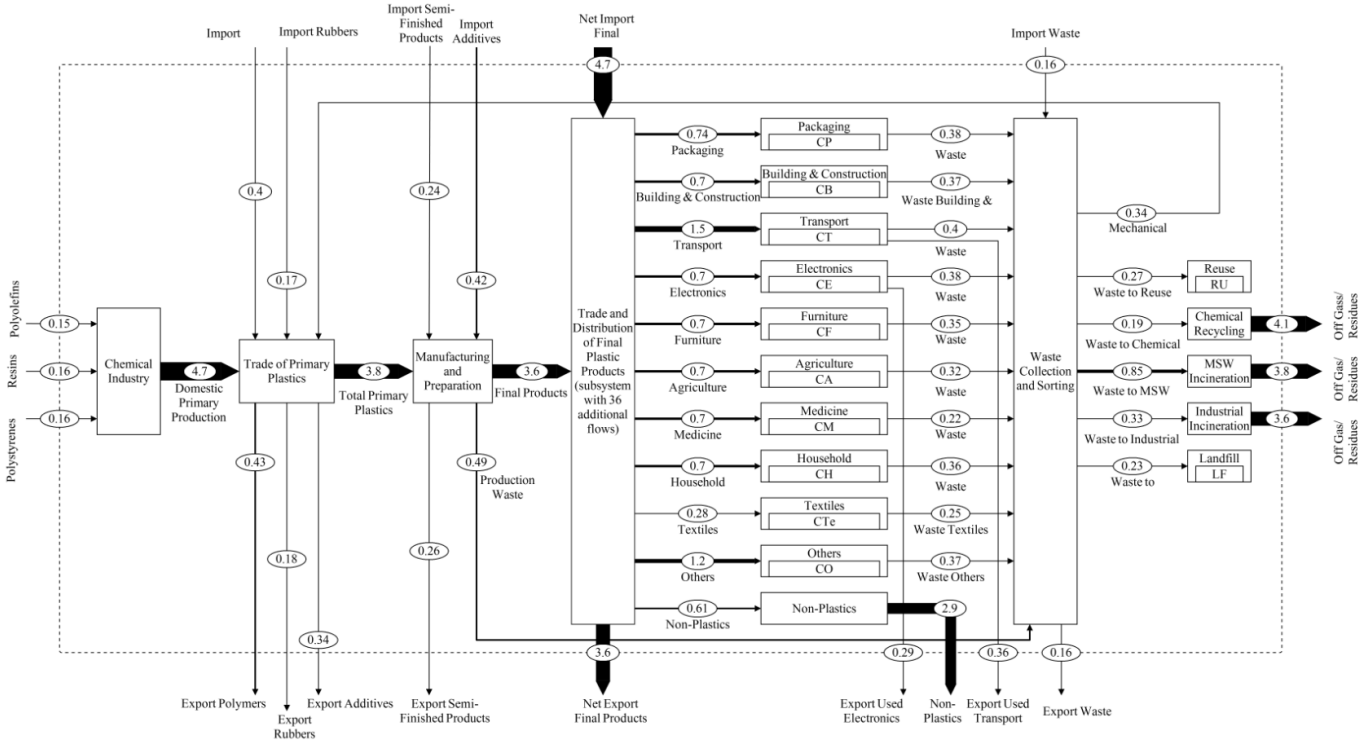


Figure A7.4: Difference between U_{ap} and U_b of the flows in the plastics MFA displayed as a flowchart, indicating the degree of data reconciliation per flow.

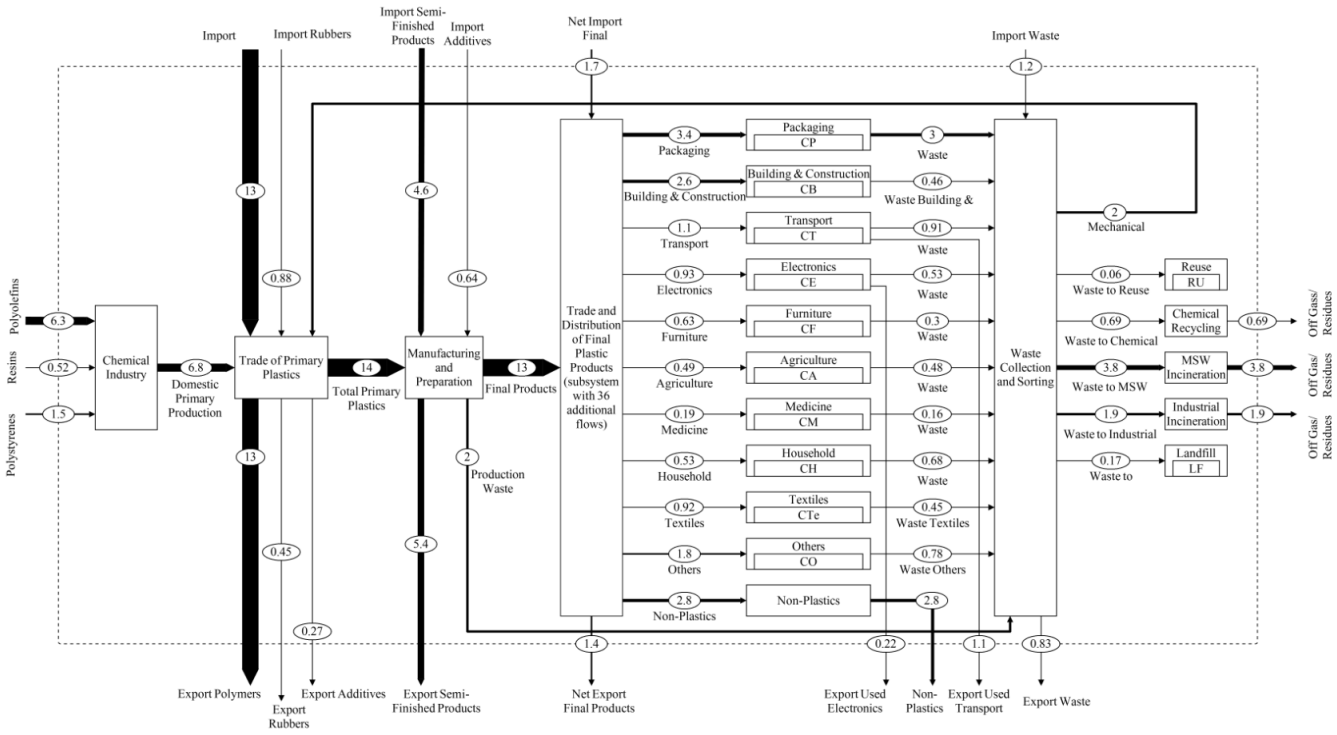


Figure A7.5: Uncertainty $U_{b,w}$ of the flows in the plastics MFA displayed as a flowchart. Flow widths are proportional to the weighted uncertainty per flow. For the total system, it is U_b=168.

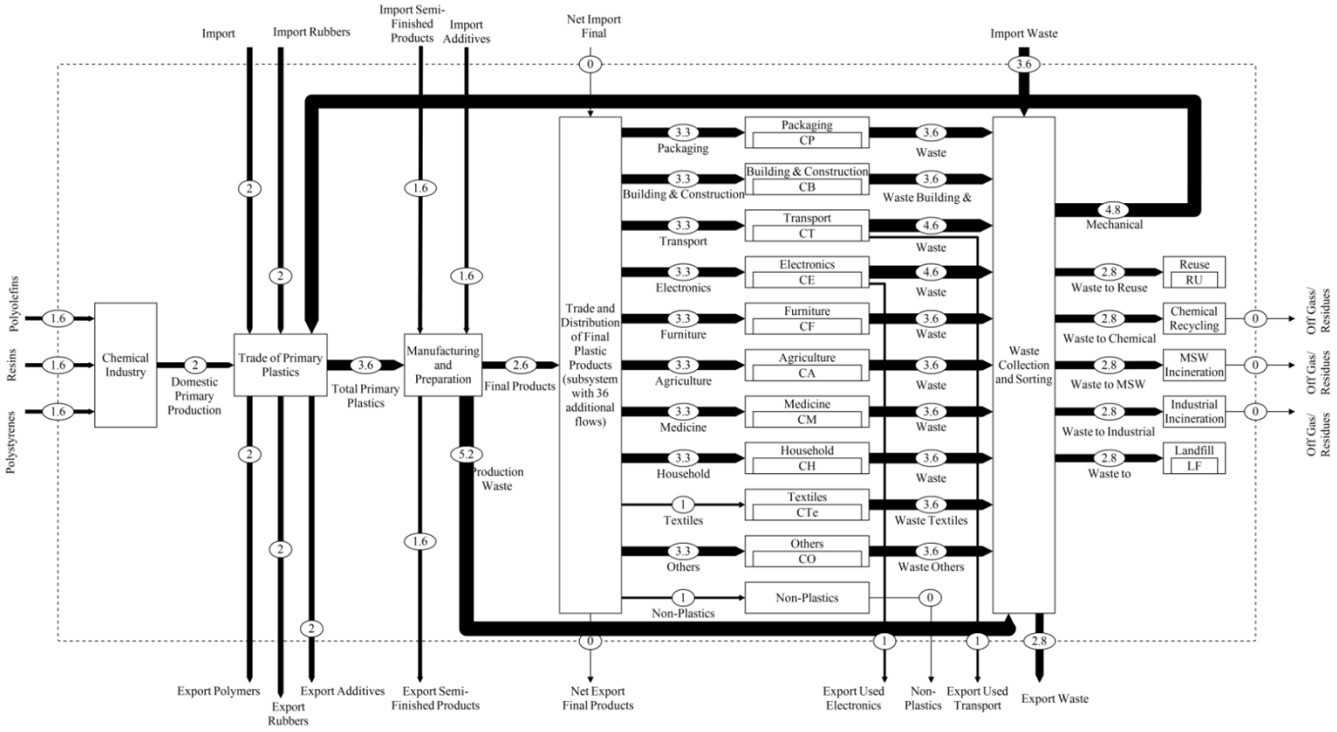


Figure A7.6: Complexity C of the flows in the plastics MFA displayed as a flowchart. Flow widths are proportional to the contribution per flow to C. For the total system, it is C=297.

Appendix 8: Copies of the three journal articles

Article I: Schwab, O., O. Zoboli, and H. Rechberger. 2016. A Data Characterization Framework for Material Flow Analysis. *Journal of Industrial Ecology*.

Article II: Schwab, O., D. Laner, and H. Rechberger. 2016. Quantitative evaluation of data quality in regional Material Flow Analysis. *Journal of Industrial Ecology*.

Article III: Schwab, O. and H. Rechberger. Information Content, Complexity and Uncertainty in Material Flow Analysis. *Journal of Industrial Ecology*. Under revision.

A Data Characterization Framework for Material Flow Analysis

Oliver Schwab, Ottavia Zoboli, and Helmut Rechberger

Keywords:

database analysis
data characterization matrix
data quality
industrial ecology
information quality
material flow analysis
national resource budget

Supporting information is available on the *JIE* Web site

Summary

The validity of material flow analyses (MFAs) depends on the available information base, that is, the quality and quantity of available data. MFA data are cross-disciplinary, can have varying formats and qualities, and originate from heterogeneous sources, such as official statistics, scientific models, or expert estimations. Statistical methods for data evaluation are most often inadequate, because MFA data are typically isolated values rather than extensive data sets. In consideration of the properties of MFA data, a data characterization framework for MFA is presented. It consists of an MFA data terminology, a data characterization matrix, and a procedure for database analysis. The framework facilitates systematic data characterization by cell-level tagging of data with data attributes. Data attributes represent data characteristics and metainformation regarding statistical properties, meaning, origination, and application of the data. The data characterization framework is illustrated in a case study of a national phosphorus budget. This work furthers understanding of the information basis of material flow systems, promotes the transparent documentation and precise communication of MFA input data, and can be the foundation for better data interpretation and comprehensive data quality evaluation.

Introduction

Material Flow Analysis and Information

Material flow analysis (MFA) is a standardized input-output system analysis methodology for the systematic investigation of material flows into, within, and out of a given system and its associated material stocks (Brunner and Rechberger 2004). It has been widely applied for the analysis of material systems in resource and waste management. MFA is commonly used for plant-level analyses or for regional analyses, such as national resource budgets, which represent a detailed balance of a national economy for a particular substance or good. There are numerous examples of static MFAs (for 1 year) and dynamic MFAs (for a time series).

Studies of anthropogenic material systems certainly reveal otherwise unknown information (Chen and Graedel 2012). Although studies of material systems can provide information, they also depend on information in their production process, and a lack of useful information can be a limiting factor to the level of detail of an analysis and its validity. Most often, the results are inherently limited in accuracy and thus in their reliability in subsequent decision-making processes (Graedel et al. 2004; Chen and Graedel 2012). The influence of information shortcomings on the feasibility and quality of MFAs have not been systematically investigated. However, if MFA is seen as a way of compiling data to create information about material stocks and flows and to aggregate this information to create knowledge about material systems, the quality and quantity of its very fundamental elements, data, is substantial. Despite its importance for the validity of modeling results, there is no collective understanding about what information, or more specifically, data, in MFA is and how it can be characterized. In a

Address correspondence to: Oliver Schwab, Institute for Water Quality, Resource and Waste Management, Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria.*Email*: oliver.schwab@tuwien.ac.at, *Web*: http://iwr.tuwien.ac.at/ressourcen/

DOI: 10.1111/jiec.12399

Editor managing review: Seiji Hashimoto

Volume 00, Number 0

wider sense, science does not have a complete idea of what information actually is, but it is stated that it is a multilayered phenomenon that is to be examined on various levels (Arndt 2004). This indicates that the elaboration of a general concept of information may not be feasible, but specific scientific formulations for certain aspects can be intended. An MFA-specific concept of quantitative information is proposed in this article.

Data in Material Flow Analysis

Data quality and data quantity are constitutive for environmental modeling, also in MFA, which is often based on cross-disciplinary data. These unstructured data can have different formats and qualities and come from heterogeneous sources, such as official trade statistics, scientific literature, consumer behavior studies, and expert estimates. In some cases, extensive statistical data, such as lab data on substance concentrations, might be available, but analysts usually have to cope with isolated values. An isolated value is a specification of an entity, such as a material quantity or a substance concentration represented not by a set of data records, but by one singular datum. Consequently, statistical methods of parameter uncertainty evaluation are often inadequate in MFA practice (Hedbrant and Sörme 2001). Additionally, relevant data may be confidential, lost, highly aggregated, or outdated, or real-world phenomena may be too complex to be directly measured and must be surveyed or derived in other indirect ways. The background of data is not always transparent because of missing metainformation, and the data can be inaccurate owing to measurement and collection errors or biased by the interests of the data producers. Data inaccuracy and unrepresentative data have been identified as two major sources of uncertainties in environmental modeling (Björklund 2002) and in MFA (Danius and Burström 2001). Recognizing the shortcomings of MFA data in combination with the mentioned variety of sources and the various ways collected data are applied in the analysis process, the databases of studies are not always comprehensible for agents other than the producer, and the systematic evaluation of data quality is limited. Though different ways of dealing with data uncertainties in MFA and related disciplines have been proposed (Laner et al. 2014; Refsgaard et al. 2007; Hedbrant and Sörme 2001; Weidema and Wesnæs 1996), a sound understanding of the nature of MFA data remains a subject for research.

In this article, a framework for consistent description and characterization of a priori MFA data (before application in a model) is presented. This can be the basis for (1) analysis of an MFA study's database structure and (2) data quality evaluation. The focus of this article is on (1) and the proposed concept is illustrated by application to a regional MFA of phosphorus. The benefits and shortcomings for MFA practice are discussed. This study is based on the idea that a sound understanding of applied data is necessary for data quality evaluation and uncertainty analysis in MFA (Laner et al. 2014), especially in the presence of scarce information and isolated values.

Material Flow Analysis Data Characterization Framework

The core of this framework is a data characterization matrix (DCM) that facilitates the systematic documentation and characterization of MFA data. Before the DCM is introduced and applied to a case study, central terms are defined.

Terminology

This terminology is to provide a conception of data and information in MFA as a basis for precise communication within and beyond the research community, and to contribute to a common understanding of quantitative information in MFA. The terminology is the foundation of the data characterization framework.

Material Flow Analysis System Elements

MFA system elements are the components of material flow systems, that is, flows, processes, stocks, and materials (Brunner and Rechberger 2004). Flows are specified as mass per time, processes as dimensionless transfer coefficients, and stocks as mass. Material is an umbrella term for goods and substances. Each system element is assigned a specific number as an identifier, that is, a flow or process number. Cross-boundary flows (flows that leave or enter the system) are called imports and exports, and flows within the system (between processes) are called internal flows. One or more related processes and associated flows can be referred to as sectors, such as industry and trade sector or consumption sector. Designating sectors can improve the comprehensibility and ease of communication about material flow systems. It also enables comparing systems that differ in their overall composition of processes, but consist of similar sectors.

Entity, Data Element, and Attribute

An entity is a real-world phenomenon or real-world object, and its realizations are represented as data. If data in MFA are considered to be quantitative information, data are representations of entities as numeric values (see Floridi 2013). That is, an entity can be represented by a data element (isolated value, interval, or data set). The number of a study's data elements can be larger than the number of entities because more than one reference could be available for quantification of an entity (e.g., three independent references on a phosphorus concentration of an agricultural good, i.e., three data elements on one entity). The total of all data elements per entity is referred to as information element. MFA data attributes are data-associated annotations concerning statistical properties, meaning, origination, and application of the data. Attributes can be designated as the "characteristics of data" (Wang et al. 1995) and specify a data element, the relation to the entity it represents, its origination and formation process, and its relation to the application context.

Information Level

Four levels of information in MFA can be distinguished (figure 1). The first information level is data element, as

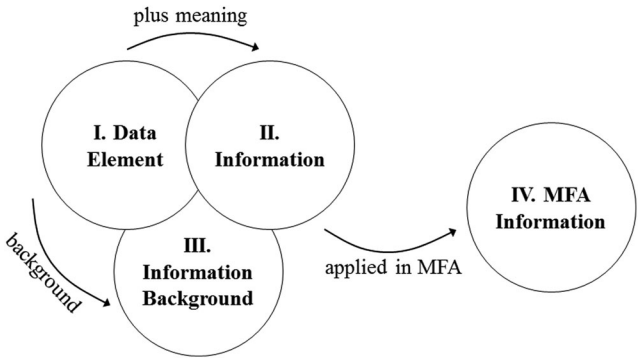


Figure 1 MFA information is information in MFA context: A data element plus meaning forms information, this information has a background, and in the context of a MFA study it forms MFA information.

described above, and a data element plus meaning forms information according to Floridi (2013). Information background represents the origination and forming process of the piece of information. Placed in context, it forms MFA information. For example, the entity "Aluminum content of a beverage can" is to be specified for an MFA study. The datum is, say, "95." This forms information, with its meaning "Aluminum content of a standard beverage can in central Europe in 2010, in %." The information background is, for example, that it has been measured by an academic research group by X-ray analysis, but the specific observation method and the number of samples are unknown. It forms MFA information when applied in an MFA study as a material specification for a designated flow or stock. MFA information can be described by sets of attributes that are arranged according to the four distinct information levels, as proposed in the below introduced data characterization matrix.

Data Semantics

Semantics refers to the intrinsic meaning of a piece of information, and data that are meaningful and truthful can become information (Floridi 2013). For example, the data at hand may describe the "phosphorus content of national annual crop production in 1990." This specification of the datas' meaning lacks semantic precision, because the notion of "crop" is ambiguous. It is not known whether it refers to food crops, to cereals, or also to energy crops and industrial crops. Data semantics can also change over time (Madnick and Zhu 2006), such as when the variety of cultivated crops changes. Unclear data semantics can lead to data misinterpretation and, consequently, to drawbacks in data quality.

System Relation and System Adequacy

System relation refers to the sphere that determines the data (such as market processes, technological state of the art, or biosphere) and the variability of a datum over time, space, and other potential relations. Other relations can be, among others, technology (e.g., productivity rates can differ between production plants) or reference units (such as data that refer to fiscal years instead of calendar years). MFA data should be adequate for the studied system with respect to time, space, and potential further relation. For example, data from a neighboring country might be temporally adequate, but spatially inadequate, or might be inadequate because they describe a different technical process (further relation).

Autonomy and Application of Data

Data that can be directly introduced in a model for the description of system elements are referred to as to be autonomous in their application. Often, there are no ready-to-apply autonomous data available for the description of MFA system elements. These need to be instead quantified by the combination of several nonautonomous data elements. Data elements can be applied in an MFA study as one of the typical utilization types (flow, flux, stock, transfer coefficient, and material). Other data elements, such as areas and numbers, are summarized as precursors. For example, readily applicable data of mineral phosphorus fertilizer use from consumption statistics, given in mass of phosphorus per year, is autonomous for the purpose of a national phosphorus budget. In contrast, the flow of phosphorus in animal manure (flow, tonnes per year) is nonautonomous if it needs to be calculated from the number of animals (precursor, dimensionless), excretion per animal type (flux, kilograms per animal per year) and the phosphorus concentration of animal excrement (material, %). The more nonautonomous data elements there are to be combined for the description of a system element, the higher is the number of potential data quality impairments.

Origination of MFA Data

Data for MFA can be acquired either from direct observations, such as measurements, monitoring, or counting ("empiricism"), or can be abstracted from given information. In contrast to empiricism, the latter is in this context referred to as "derived" and is divided into three categories: "mainly from data" (such as reporting data that is aggregated by statistical offices); "mainly from assumptions" (such as data from models with many assumptions because of a scarce database); and "from speculation" (such as guesses).

Variety and Disparity

The attributes variety and disparity describe the complexity of a population. Variety refers to the number of potential real-world objects an entity refers to, disparity to the spread of these real-world objects' realizations. For instance, "copper content of smartphones" can refer to a vast number of different smartphones (high variety), and the copper content of these smartphones can span a wide concentration range (high disparity). In contrast, both the variety and disparity of the "aluminum content of aluminum cans" are comparably small, because the number of different types of aluminum cans is limited and the range of the aluminum content is rather narrow (between 95% and 99%). A more precise specification of a data element's meaning (e.g., to a particular type of smartphone) can reduce variety and disparity.

Table 1 Structure of the data characterization matrix by information levels and attribute groups

Information level	*Attribute group*	*Description (no. of attributes)*	*Attributes*
Data element	Statistical characteristics	Documentation of statistical information on a data element (10)	Data element form, location parameter, value (numeric), n, minimum, maximum, distribution (form), distribution (parameter), dispersion (measure), dispersion (numeric)
Information	Semantics	Specification of the meaning of a data element (2)	Description of meaning, semantic precision
	Scale	Specification of the format of an entity (8)	Entity category, entity class, unit, sphere, property type, mathematical form, minimum (potential), maximum (potential)
	Complexity	Description of the complexity of an entity (2)	Variety, disparity
Information background	Availability	Distinction if wanted information does exist and is accessible or not (3)	Existence, accessibility, access restriction
	Communication	Documentation of how a piece of information is communicated (3)	Communication type, access type, frequency
	Producer	Documentation of the agent that produced the piece of information, for example, an authority (3)	Producer category, producer type, reference
	Origination	Documentation of the data collection method, for example counting or industrial monitoring (3)	Origination category, origination type, origination type quality
MFA information	Application in MFA	Description of how a piece of information is applied in the MFA study (4)	Utilization type, autonomy, layer, type of good
	System relation	Description of the relation between a piece of information and the studied system (6)	Primary determination, temporal variability, trend, spatial variability, further relation, variability by further relation
	System adequacy	Description of a piece of information's adequacy to (respective divergence from) the studied system (5)	Temporal divergence, spatial divergence, further divergence, adaptation (type), adaptation (quality)

Note: A more detailed description of the data attributes is provided in appendix 1 in the supporting information on the Web.

It is to be considered that data quality can decrease because of improperly understood data semantics and limited context knowledge (Madnick and Zhu 2006) and that information always has a subjective element (Arndt 2004). This is considered in the data characterization framework, which, at the same time, is designed for a high degree of transparency and replicability. Key to the framework is the characterization of MFA data by specification of data attributes in a DCM.

MFA Data Characterization Matrix

The database of a material flow system is documented, structured, and analyzed in the DCM. The DCM has been developed in an iterative process by the analysis of several regional MFAs (Schwab and Rechberger 2014a). In the matrix, 49 data attributes are assigned to each data element of a study. The DCM is structured according to the four information levels (figure 1), and related attributes are grouped in attribute groups (table 1). A more detailed description of each data attribute is provided in appendix 1 in the supporting information on the Journal's website. For application of the DCM to a given MFA database, each of these data attributes is specified individually. For specification of the attributes, a code has been developed. By this code, attributes are assigned to particular measurement scales (absolute, nominal, binary, or ordinal) and ranges of possible data attribute specifications are provided. This facilitates the consistent completion of the matrix, also when applied by different researchers to different regional MFAs, and enables automated analysis of a DCM once completed. The DCM code is provided in appendix 2 in the supporting information on the Web and an example of a completed matrix in appendix 3 in the supporting information on the Web.

In the following, the data characterization framework is illustrated in a database analysis of a national resource budget. This database analysis consists of three steps, which are (1) creation of data inventory, (2) characterization of data elements, and (3) analysis of data attributes. In (1), all system elements and the respective data elements are listed in the DCM. In (2),

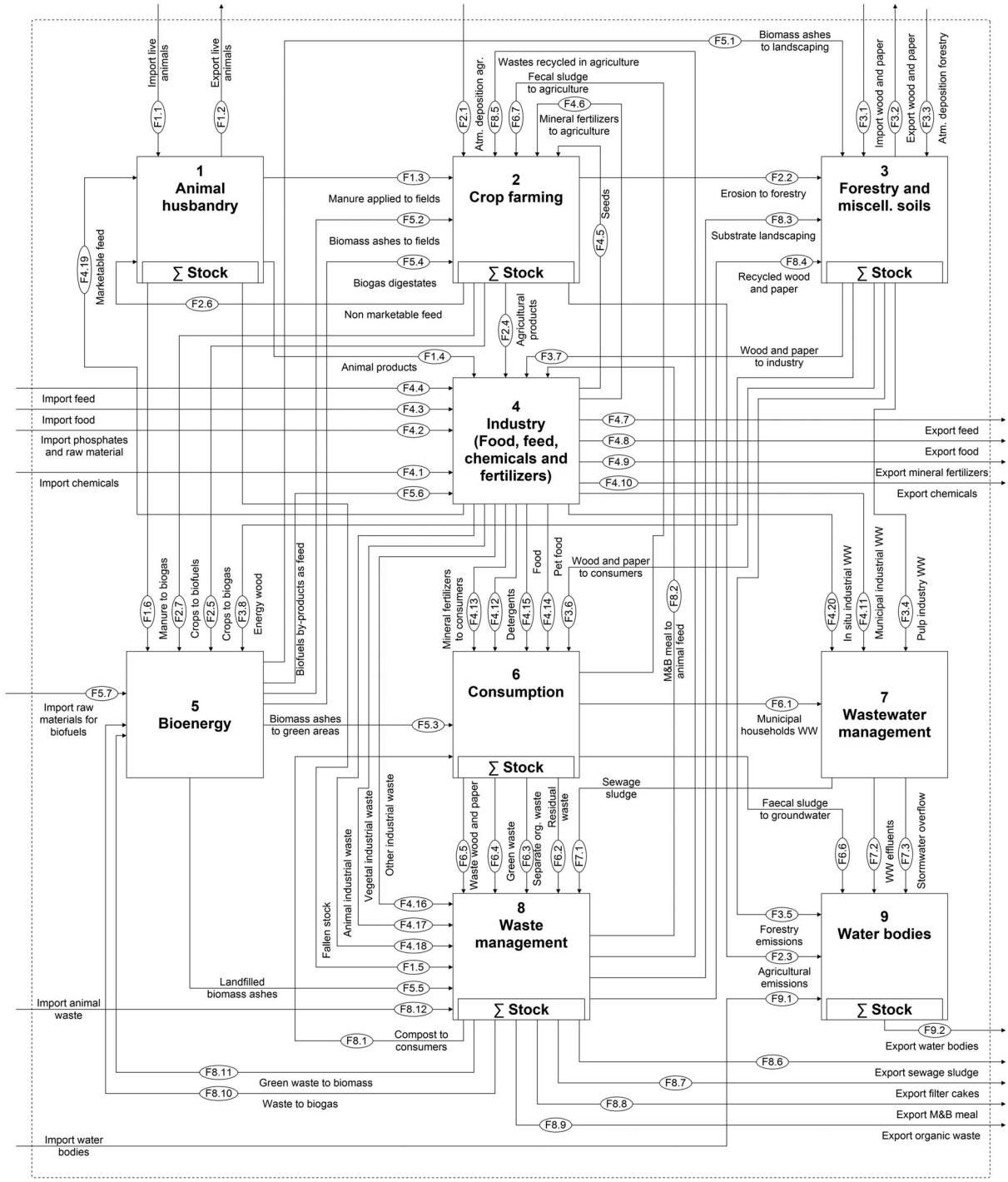


Figure 2 Schema of the Austrian phosphorus budget according to Zoboli and colleagues (2015).

the attributes are specified with the help of the DCM code, and in (3), the DCM is analyzed attribute wise.

Application of the Data Characterization Framework

The data characterization framework is applied to the 2009 national phosphorus (P) budget of Austria (figure 2) (Zoboli et al. 2015), which is based on the work of Egle and colleagues (2014). A comparatively sound database for quantification of material flows and stocks of this phosphorus budget is available. Data uncertainties were assessed by an approach by Laner and colleagues (Forthcoming) and range from 10% to 90%. Nine out of ten flows have less than 40% uncertainty, and two thirds of the flows have less than 30%. These relatively low uncertainties (compared with other regional MFAs) underline the

Table 2 Key data on the complexity and the data basis of the analyzed national resource budget (phosphorus in Austria 2009)

Database characteristic	*Quantity*
Number of flows in main system	72
Number of processes	9
Number of subsystems	8
Number of stocks	7
Total number of collected data elements	308
Total number of entities	172
Average entities per flow	2.4
Average data elements per entity	1.8
Share of flows that can be described directly by autonomous data (%)	20
Isolated values (%)	75

database's robustness. Key information on the system and the applied database is provided in table 2. The phosphorus budget of Zoboli and colleagues is a flow-based model in which the number of applied transfer coefficients is kept to a minimum. Respectively, the scope of the here presented case study is limited to the evaluation of these phosphorus flows in the main system and does not include processes and subsystems.

Data Inventory

The total of 308 data elements and all assigned data attributes are inventoried in the DCM (table 1 and appendix 3 in the supporting information on the Web). As listed in table 2, these 308 data elements are used for the description of 172 entities and are aggregated for the description of 72 flows. Twenty percent of these flows are quantified directly by autonomous data and 80% by the combination of data on two or more entities.

Characterization of Data Elements and Analysis of Data Attributes

The elements of the data inventory are evaluated by specification of data attributes according to the code for data characterization (appendix 2 in the supporting information on the Web). For a complete DCM of the case study and detailed specification of data attributes for all data elements, see appendix 3 in the supporting information on the Web. Exemplarily, selected attributes are analyzed in the following: data producer (figure 3a), data origination (figure 3b), utilization type (figure 4a), entity class (figure 4b), type of good (figure 5), and primary determination (figure 6). Please note that the quantities given here are not material quantities, but information quantities. The number of samples *n* in figures 3, 4, 5, and 6 relates to the number of collected data elements (n = 308) or the number of entities (n = 172).

More than half of the data were collected from authorities, around 40% from scientific sources (figure 3a). Generally speaking, data on material flows stem from authorities and data on material qualities (composition) from science. Approximately 40% of the data elements are from empirical collections (such as measurement or counting) and 55% are derived (either from data, assumptions, or speculations). Most prominent are reported data from third parties that are aggregated by authorities (figure 3b), such as official trade statistics. These contribute to the generally more robust database for cross-boundary flows (e.g., for imports of goods) in contrast to the often weaker database within the system (e.g., in the consumption sector).

The two most prominent references, namely, reporting data from statistical offices and empirical data from scientific measurements, are complemented by data from additional sources. Expert estimations are important especially in the consumption and waste management sectors, assumptions in the bioenergy sector, and scientific models in the waste management and crop farming sectors. For animal husbandry, simple calculations based on data from authorities and science complement directly applicable data. More than 40% of the data are communicated in reports, 35% in online databases, and 10% in scientific journals or books (cf. attribute no. a305 [access type] in appendix 3 in the supporting information on the Web). The number of data elements per entity (on average, 1.8; see table 2) is less than or equal to four in 95% of the cases, and 75% are isolated values.

Most of the collected data describe material flows (figure 4a) and come in the format "mass/time" (figure 4b). Approximately one sixth of the collected data are precursors, mainly on numbers and areas, and need to be combined with other data before introduction to a model.

Forty percent of the collected data describe waste, 25% consumer goods, and 20% industrial goods (figure 5). The label "none" refers to other entities, such as conversion factors or areas. Although the waste management sector has less flows than other sectors, such as industry (see figure 2), most of the collected data are on waste. This indicates that in this case, less directly applicable, autonomous data for the description of the waste management sector is available and, in consequence, the overall data search effort is greater.

The attribute "primary determination" (figure 6) refers to the spheres that primarily determine the data values. For example, the phosphorus concentration of common wheat is primarily determined by the biosphere, and phosphorus removal rate of a sewage plant by the applied technology. In the analyzed study, 40% of the data elements are primarily determined by market activities (such as domestic production of agricultural goods), 10% by technology (e.g., phosphorus removal rate from wastewater), 8% within the sociosphere (e.g., consumer behavior), 30% in the biosphere (such as phosphorus content of crops), and 6% in the geosphere (such as discharge of rivers per time unit). Data that are primarily determined by political decisions are applied mainly in the waste management and bioenergy sectors (e.g., amount of phosphorus in fecal sludge applied on agricultural fields). Examples of applied scientific rationales are molecular masses of phosphorus and phosphorus compounds.

Although the primary determination of data is not always unequivocal, it can indicate the main factors that shape the

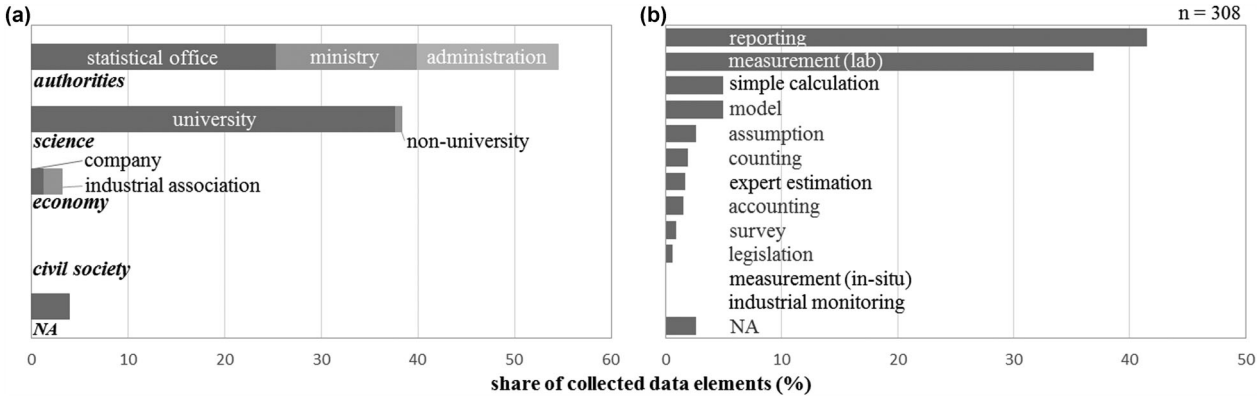


Figure 3 Data in the phosphorus case study: (a) producer category and producer type and (b) origination type. NA = not available.

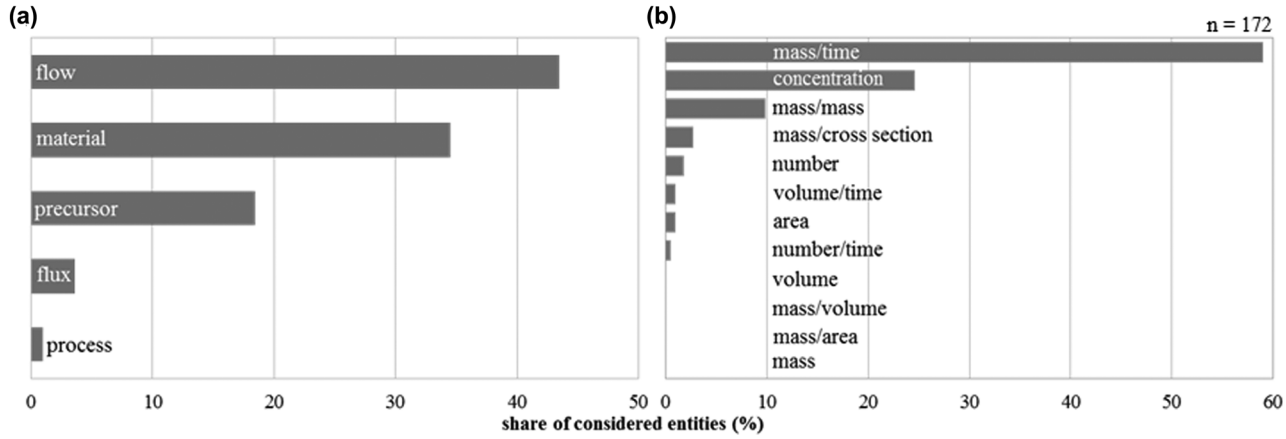


Figure 4 (a) Utilization describes the utilization type of data used in the study, and (b) entity class of data describes the format of the collected data. Concentration is in mass-%; mass/mass refers to other entities such as productivity rates.

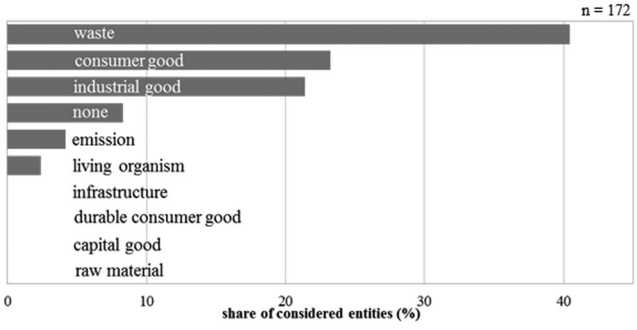


Figure 5 Collected data relating to different types of goods. None = no goods but other entities, such as areas or conversion factors.

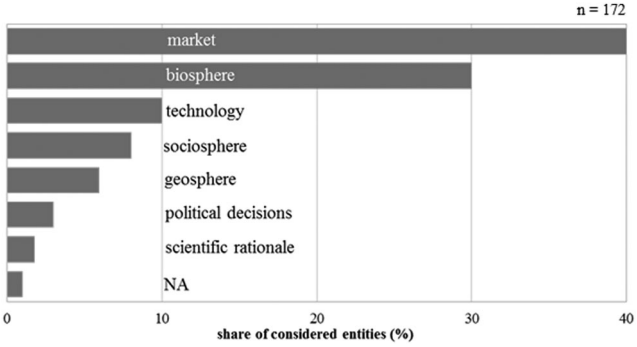


Figure 6 Primary determination (mechanisms that primarily determine the value of data and their change over time) of data within the anthroposphere and the natural environment. NA = not available.

data of an MFA and thus the material system itself. Clearly, the quantity of applied data elements does not necessarily correlate with the physical quantity of the material flows of a study. However, this can contribute to the identification of a system's main driving phenomena. The driving factors of material systems in terms of physical quantities have been investigated by Klee and Graedel (2004). Transferring this idea from physical quantities to information quantities, the DCM can be used to reveal the driving mechanisms of a material system from its database. The database structure of the case study indicates that, regarding its information quantities, the Austrian P budget appears to be strongly perturbed by anthropogenic activities, but to be not entirely dominated because there is still a prominent influence of the natural environment (i.e., biosphere and geosphere; see figure 6) on the material system.

Discussion

The data characterization framework can contribute to MFA in practice. First, the data terminology enables more precise communication within and beyond the research community. The DCM facilitates the systematic documentation of MFA data and designated attributes. Attribute-wise data evaluation draws a compact picture of this database as illustrated in figures 3, 4, 5, and 6. This promotes a simple and condensed representation of MFA databases, for instance, in reports or publications, as an alternative to communicating extensive data tables. Systematic tagging of MFA data with attributes can further the understanding of the information basis of a study, enables comparing different MFA studies to one another, and can give indications regarding the quality of the data.

Nevertheless, it has to be considered that the database is subject to the scope and level of detail of a study, which is determined by the focus of the research. In the analyzed P budget, processes such as crop farming are rather treated as "black boxes," whereas wastewater and waste management processes were ranked higher in the specific interest of the research group and were thus studied in more detail. From the experience of this study, it can be said that the quality of available information decreases when moving downstream the material flows (see also Mao et al. 2008; Graedel et al. 2004). Previous database analyses indicated the tendency of decreasing data quality when moving from the main system into specific subsystems with higher level of detail (Schwab and Rechberger 2014b). Both tendencies are also owing to a decreasing share of "hard data" from authorities and science, an increasing share of speculations and expert estimations, and the decreasing autonomy of the data.

Attribute specifications are, in part, subject to authors' judgments. Therefore, these specifications can be argued over by third persons with diverging perspectives. Author judgments on MFA data have also been an element of previous studies (Graedel et al. 2004; Hedbrant and Sörme 2001). The novelty of the approach presented here is that it is not the data as such that are judged by the authors, but individual data characteristics (i.e., data attributes). Inevitably, subjectivity is intrinsic to these judgments. Nevertheless, subjectivity is also intrinsic to information per se. Ignoring the subjective part of information can restrain a comprehensive understanding of it (Arndt 2004; Berger and Berry 1988). Therefore, subjectivity is also an element of the here-presented framework, even if it is controlled by a standardized and transparent procedure, which, however, can facilitate a discourse about MFA databases and collective learning of material systems.

In the case study, metainformation on the meaning and the formation process of data was sometimes found to be limited, although it is imperative for data producers and data publishers to provide this information. Based on the here-presented database analysis, it can be said that this information may be lost or become imprecise in the scientific publication and citing process. Over time, data can appear "just to be there," without precise knowledge about its initial meaning and collection method, which might lead to poor data quality estimations and poor application of the data. Moreover, data can be misinterpreted not only because of ambiguous data semantics, but also because of diverging reference units. This was found to be the case in the P budget, because most of the data refer to calendar years (1 January–31 December), but some do refer to fiscal years (often from June to June).

This framework can be enhanced with further experience in MFA database analysis. It is not recommended that the DCM is applied posterior to an analysis, but rather simultaneously with the data collection process. The net working time for a database analysis of a study with the extent of the above-described P budget is approximately 60 to 80 hours.

Conclusion and Outlook

MFAs are typically based on diverse and often scarce information. Most often, it is not practicable to evaluate the overall quality and compare one MFA to another. Statistical methods are not always sufficient for the characterization of MFA data. Alternative data characterization methods that consider the inherent subjective notion of information and metainformation, such as the framework presented here, can complement existing practice. The framework consists of a data terminology and a characterization matrix for MFA data. It facilitates the systematic characterization and communication of databases. This is a step toward a comprehensive understanding of the nature and role of quantitative information in MFA. It can contribute to data quality evaluation and high-quality MFAs, and it can enable a comparison of analyses and their databases to another.

As indicated by the P case study, statistical offices (aggregated reporting data on material quantities) and scientific literature (measurement data on material qualities) are the central data sources. Data from industry and also from civil society (e.g., from interest groups) can be more relevant for similar studies of different substances. The database for cross-boundary flows is found to be better than for flows within the system. This is especially because of detailed official foreign trade statistics. In contrast, institutionalized statistics such as the latter are limited within the system, for example, in the consumption sector. The results indicate that there is a general tendency of the databases to become weaker from upstream to downstream sectors, that is, from primary production and industry to waste management. Especially in waste management and in consumption, data producers are required to be more active in providing disaggregated and transparent data in consistent formats.

The assignment of data attributes to MFA data is sometimes more ambitious than it may seem at first. Metainformation is often of limited availability, because it is often unpublished, has become lost, or has become imprecise over time, and its retrieval can be complex. It is recommended that data users and producers communicate and document the background information of data as precisely as feasible. This can be performed by means of this data characterization framework. It can then be the basis of further research toward the systematic evaluation of MFA data quality. For a comprehensive MFA data

uncertainty analysis that considers both epistemic uncertainty (owing to lack of knowledge) and aleatory uncertainty (owing to natural variability) (Dubois and Guyonnet 2011; Clavreul et al. 2013; Laner et al. 2015), a profound understanding of the characteristics and meaning of MFA data is imperative. The here-proposed terminology and procedure can be the basis of coherent MFA data communication, better data interpretation, and attribute-based data quality evaluation across different studies and research groups.

Acknowledgments

This research was funded by the Austrian Federal Ministry of Science, Research and Economy. The authors thank the colleagues at the Christian Doppler Laboratory for Anthropogenic Resources at Vienna University of Technology for their support, and three anonymous reviewers for their comments on a previous version of this article.

References

Arndt, C. 2004. *Information measures: Information and its description in science and engineering*. Heidelberg, Germany: Springer.

Berger, J. O. and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76(2): 159–165.

Björklund, A. E. 2002. Survey of approaches to improve reliability in LCA. *The International Journal of Life Cycle Assessment* 7(2): 64–72.

Brunner, P. H. and H. Rechberger. 2004. *Practical handbook of material flow analysis*. Boca Raton, FL, USA: Lewis.

Chen, W.-Q. and T. E. Graedel. 2012. Anthropogenic cycles of the elements: A critical review. *Environmental Science & Technology* 46(16): 8574–8586.

Clavreul, J., D. Guyonnet, D. Tonini, and T. Christensen. 2013. Stochastic and epistemic uncertainty propagation in LCA. *The International Journal of Life Cycle Assessment* 18(7): 1393–1403.

Danius, L. and F. Burström . 2001. Regional material flow analysis and data uncertainties: Can the results be trusted? In *Sustainability in the information society. Part 2: Methods/workshop paper*, edited by L. M. Hilti and P.W. Giligen. Marburg, Germany: Metropolis Verlag.

Dubois, D. and D. Guyonnet. 2011. Risk-informed decision-making in the presence of epistemic uncertainty. *International Journal of General Systems* 40(2): 145–167.

Egle, L., O. Zoboli, S. Thaler, H. Rechberger, and M. Zessner. 2014. The Austrian P budget as a basis for resource optimization. *Resources, Conservation and Recycling* 83: 152–162.

Floridi, L. 2013. *The philosophy of information*. Oxford, UK: Oxford University Press.

Graedel, T. E., D. van Beers, M. Bertram, K. Fuse, R. B. Gordon, A. Gritsinin, A. Kapur, et al. 2004. Multilevel cycle of anthropogenic copper. *Environmental Science & Technology* 38(4): 1242–1252.

Hedbrant, J. and L. Sörme. 2001. Data vagueness and uncertainties in urban heavy-metal data collection. *Water, Air and Soil Pollution: Focus* 1(3–4): 43–53.

Klee, R. J. and T. E. Graedel. 2004. Elemental cycles: A status report on human or natural dominance. *Annual Review of Environment and Resources* 29(1): 69–107.

Laner, D., J. Feketitsch, H. Rechberger, and J. Fellner. Forthcoming. A novel approach to characterize data uncertainty in MFA and its application to plastic flows in Austria. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12326.

Laner, D., H. Rechberger, and T. Astrup. 2014. Systematic evaluation of uncertainty in regional material flow analysis. *Journal of Industrial Ecology* 18(6): 859–870.

Laner, D., H. Rechberger, and T. Astrup. 2015. Applying fuzzy and probabilistic uncertainty concepts to the material flow analysis of palladium in Austria. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12235.

Madnick, S. and H. Zhu. 2006. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering* 59(2): 460–475.

Mao, J. S., J. Dong, and T. E. Graedel. 2008. The multilevel cycle of anthropogenic lead: I. Methodology. *Resources, Conservation and Recycling* 52(8–9): 1058–1064.

Refsgaard, J. C., J. P. van der Sluijs, A. L. Højberg, and P. A. Vanrolleghem. 2007. Uncertainty in the environmental modelling process—A framework and guidance. *Environmental Modelling & Software* 22(11): 1543–1556.

Schwab, O. and H. Rechberger. 2014a. *Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen—Analyse der Daten und Quellen* [Investigation of the data requirements of national resource budgets—Analysis of data and data sources, in German]. Vienna: Vienna University of Technology.

Schwab, O. and H. Rechberger. 2014b. *Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen—Ermittlung der Datenlücken* [Investigation of the data requirements of national resource budgets—Analysis of data gaps, in German]. Vienna: Vienna University of Technology.

Wang, R. Y., M. P. Reddy, and H. B. Kon. 1995. Toward quality data: An attribute-based approach. *Decision Support Systems* 13(3–4): 349–372.

Weidema, B. P. and M. S. Wesnæs. 1996. Data quality management for life cycle inventories—An example of using data quality indicators. *Journal of Cleaner Production* 4(3–4): 167–174.

Zoboli, O., D. Laner, M. Zessner, and H. Rechberger. 2015. Added value of time series in MFA: The Austrian phosphorus budget from 1990 to 2011. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12381.

About the Authors

Oliver Schwab is a research associate at the Institute for Water Quality, Resource and Waste Management of Vienna University of Technology in Vienna, Austria. **Ottavia Zoboli** is a research associate at the Center for Water Resource Systems at Vienna University of Technology. **Helmut Rechberger** is a professor for resource management at the Institute for Water Quality, Resource and Waste Management of Vienna University of Technology in Vienna, Austria.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Supporting Information S1: This supporting information provides the attributes and codes of the data characterization matrix and a characterized database of the Austrian phosphorus budget 2009 (laid out as a large table for which digital inspection is recommended).

Quantitative Evaluation of Data Quality in Regional Material Flow Analysis

Oliver Schwab, David Laner, and Helmut Rechberger

Keywords:

data characterization
data quality
industrial ecology
information defects
material flow analysis (MFA)
substance flow analysis (SFA)

Supporting information is linked to this article on the *JIE* website

Summary

A method for quantitative evaluation of data quality in regional material flow analysis (MFA) is presented. The principal idea is that data quality is a multidimensional problem that cannot be judged by individual characteristics such as the data source, given that data from official statistics may not be *per se* of good quality and expert estimations may not be *per se* of bad quality, respectively. It appears that MFA data are never totally accurate and may have certain defects that impair the quality of the data in more than one dimension. The concept of MFA information defects is introduced, and these information defects are mathematically formalized as functions of data characteristics. They are quantified on a scale from 0 (no information defect) to 1 (maximum information defect). The proposed method is illustrated in a case study on palladium flows in Austria. A quantitative evaluation of data quality provides opportunities for understanding and assessing MFA results, their *a priori* information basis, their reliability in decision making, and data uncertainties. It is a formal step toward better reproducibility and more transparency in MFA.

Introduction

The available information base is critical for the validity of material flow analyses (MFAs) and can differ significantly among MFAs, depending also on the studied material and scope of a study. A frequently performed type of regional MFAs are so-called national resource budgets. These are detailed studies of the supply, consumption, and disposal of a specific material by national economies within a defined time period, usually 1 calendar year (see, e.g., Egle et al. 2014; Bonnin et al. 2013). The data basis of a detailed study usually includes not only black-box material flow accounts, but also more-specific information on material flows within an economy. The realizable resolution of national resource budgets (i.e., their level of detail) depends mainly on the goal of the MFA and on the available data basis (see, e.g., the comparison of Danish and Austrian phosphorus balances in Klinglmair et al. [2016]). Such data are typically unstructured, cross-disciplinary, have different formats and qualities, and come from heterogeneous sources, such as official trade tables, scientific measurements and models, industries, or associations. This implies that the databases of regional MFAs are usually highly heterogenic. In many cases, MFA data are not based on empirically well-founded data sets, but on individual, isolated values, which are not always provided in consistent formats. The available information basis is often considerably limited.

Evaluation of data uncertainties should be a component of every MFA (Rechberger et al. 2014). In regional MFAs with precisely defined temporal and spatial system boundaries, data uncertainty arises mainly from limited knowledge ("epistemic uncertainty"; see, e.g., Gottschalk et al. [2010] and Laner et al. [2014]). Natural variability ("aleatory variability") is virtually excluded. That is because for every flow, there is only one correct value, which is most probably not known with absolute certainty. Data uncertainty in regional MFAs appears to be an epistemic phenomenon. Thus, it is closely linked to the quality of the data applied. "Data quality" refers to the perceived degree of credibility of given data and the degree of belief that

Address correspondence to: Oliver Schwab, Institute for Water Quality, Resource and Waste Management, Technische Universität Wien, Karlsplatz 13, 1040 Vienna, Austria. *Email*: oliver.schwab@tuwien.ac.at, *Web*: http://iwr.tuwien.ac.at/en/resources/home/

DOI: 10.1111/jiec.12490

Editor managing review: Reinout Heijungs

Volume 00, Number 0

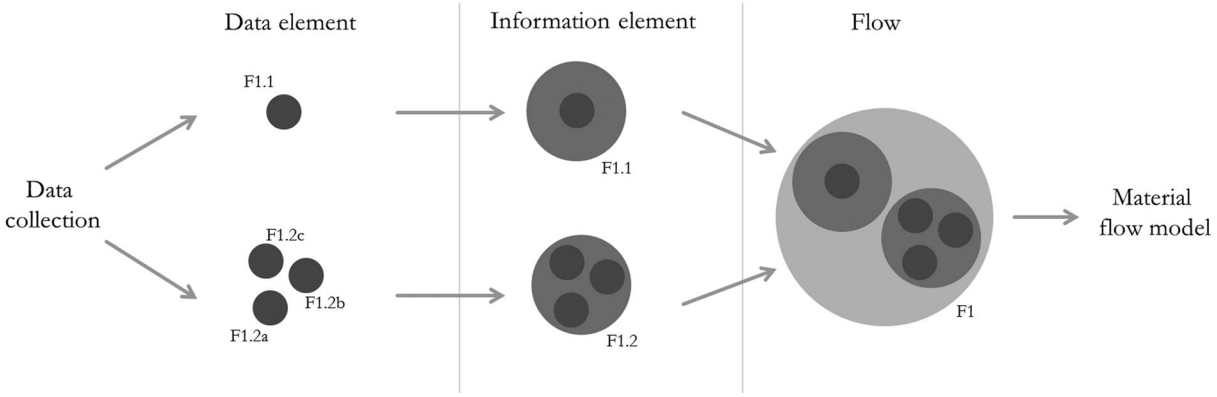


Figure 1 Illustration of a typical quantification of a material flow "F1." F1 consists of two information elements, which itself consist of one (F1.1) or more (F1.2) data elements.

the data are true in a particular context. It is, other than uncertainty, initially not a quantitative measure, but a qualitative description, such as "good" or "bad," "better than," or "worse than."

In MFA uncertainty evaluation, the heterogeneity of data is a major difficulty. Simple categorization of data quality by single data characteristics is limited, for example, because there is no apparent quality difference between data sources (Nakajima et al. 2013). Rather, data quality can be seen as a multidimensional problem that depends, for example, on the type of the quantified entity, data origination, and application context of the data. Although important methodological developments for treatment of uncertainties in MFA have been made (see, among others, Hedbrant and Sörme 2001; Gottschalk et al. 2010; Laner et al. 2014; Wu et al. 2014; Patrício et al. 2015) and methods for data reconciliation have been introduced (e.g., Cencic and Frühwirth 2015; Kopec et al. 2015; Dubois et al. 2014), methods for evaluation of data quality, a prerequisite of data uncertainty evaluation, are rare. Existing approaches to data quality in MFAs have been reviewed by Laner and colleagues (2015b), and a novel approach for calculation of data uncertainties based on five data quality indicators has been proposed.

Building on that idea, this study provides a formal procedure for systematic and quantitative evaluation of data quality based on a diversified set of data characteristics. It enables expressing data quality as a function of these data characteristics, some of which are possibly related to others. As a result, each flow of a material system can be described by one value that indicates the quality of the information applied. This value is calculated based on evaluation of all data elements applied for quantification of a particular flow. It is thus specific to any material flow of interest within a temporally and spatially defined system. For illustration, the evaluation procedure is explained in detail for one flow of the Austrian palladium (Pd) MFA (Laner et al. 2015a). Consequently, the procedure is applied to all flows of the Pd MFA and its information basis is evaluated. The benefits and shortcomings of the presented procedure are discussed.

Characterization of Material Flow Analysis Data

The data quality evaluation is based on a MFA data characterization framework, which has been proposed by Schwab and colleagues (2016). A "data characterization matrix" facilitates a structured inventory of all data applied in a regional MFA. The data in this inventory are then characterized by so-called data attributes according to a predefined syntax and scheme. MFA data attributes are data-associated annotations concerning statistical properties, meaning, origination, and application of the data. In the previous article, the data characterization matrix has been analyzed per data attribute for understanding the overall database of an MFA. In this article, the matrix is analyzed per piece of information to understand the quality of data applied in an MFA.

Three important components of the information inventory are "data element," "information element," and "flow." A substance flow is typically quantified by multiplication of two information elements ("amount of good x per time (mass/time)" and "substance concentration in good x (%)"). An information element itself can be specified by one or more data elements, where data elements are representations of real-world objects or real-world phenomena ("entities") as numerical values (isolated values, intervals, or data sets). A typical case is illustrated in figure 1, where a flow F1 is quantified by two information elements. These information elements may be quantified by data from one reference (F1.1) or by data from more than one independent references (F1.2). For example, F1.1 may be information on the number of cars imported into an economy, provided by official trade statistics. F1.2 may be information on their Pd concentration in %, provided as an expert estimation (F1.2a) or in scientific literature (F1.2b and F1.2c). The data quality of flow F1 depends on the data quality of all data elements used for its quantification.

Approach and Conceptualization

Uncertainty in regional MFA is rather an epistemic than an aleatory phenomenon, that is, not a consequence of natural

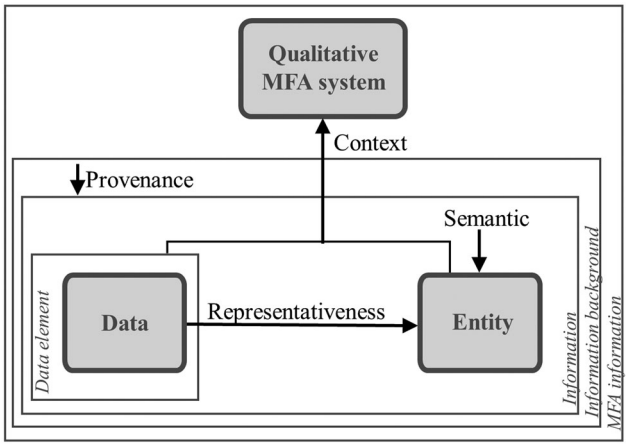


Figure 2 Concept of MFA information defects and their position in the data characterization framework by Schwab and colleagues (2016). "Data" are numerical values, "entity" is a real-world object or phenomenon described by an information element, "qualitative MFA system" is a system to be quantified by introduction of quantitative information. MFA = material flow analysis.

variability, but of imperfect knowledge. Knowledge shortcomings are here expressed as "defects of information" (Dubois and Prade 2010). Information defects indicate the deviation of given information from a desired state of perfect knowledge. They are expressed on an ordinal scale from 0 (no information defect) to 1 (maximum information defect). The four information defects, "semantic," "representativeness," "provenance," and "context" (ID_S, ID_R, ID_P, and ID_C), appear to be relevant for regional MFAs (figure 2).

ID_S refers to the semantic precision, or respectively, imprecision of the meaning of data (Does the specification "smart phones" also refer to mobile phones from before the technological leap, which are still "out there"?). ID_R indicates how well a given data element represents the entity of interest (Is the complex entity "Pd concentration of mobile phones" quantified based on one or more measurements or independent references?). ID_P considers the origination and collection method of a data element (How reliable are the information producer and the data collection method?). ID_C designates how well a given data element fits the context of a study (Is the data element timely and does it refer to the geographical area studied?). These information defects are, to some degree, similar to data quality indicators found to be useful in previous studies (Laner et al. 2015b) in terms of, for example, the correlation in the dimensions "time," "space," and "further," which are here part of the context information defect ID_C. Other defects, such as the semantic information defect ID_S or the representativeness information defect ID_R, are new concepts of the approach presented here.

The approach is illustrated in detail for flow ten (F10) of the Pd case study (see paragraph "Case study on Pd flows in Austria 2011" in Laner et al. [2015a]). A fully characterized information inventory of the Pd flow study is provided in supporting information S1 available on the Journal's website. Flow F10 refers to the Pd flow in flat screens sold in Austria in 2011. Laner and colleagues quantified flow F10 based on two information elements. First, this is the per capita flow of flat screens (information element F10.1) and, second, the Pd content in flat screens (information element F10.2). F10.1 was calculated from data on the 2010 German market and related assumptions. For F10.2, information from a scientific report providing German data of the year 2010 was used.

The information defects illustrated in figure 2 are exemplarily qualified for information element F10.2 (Pd content in flat screens). The information element F10.2 has a semantic information defect (ID_S) because "flat screens" is not a clear specification given that there are different types of flat screens. F10.2 refers to a complex entity given that different types of flat screens (attribute "variety") may also differ in their Pd content (attribute "disparity"). This complex entity is quantified based on one reference. A complex entity, in combination with a small number of references or samples, induces a representativeness information defect ID_R. Because no information is provided on the data collection method, there may also be a provenance information defect (ID_P). The data do not fit the actual system context (Austria 2011) because they are for Germany in 2010. Consequently, there is a context information defect ID_C.

This vague qualitative description of information defects enables first estimates on the overall quality of the data. A formal procedure for quantitative estimation of the qualitatively introduced concept of information defects is proposed in the following.

Formalization

The information defect per flow ID_F is quantified in three steps. First, the quality of each data element is described by a set of four defects ID_i (ID_S, ID_R, ID_P, and ID_C; see figure 2). Second, each information element is described by one total information defect (ID_{tot}), which is an aggregation of the ID_i of the respective data element or data elements. Third, the data quality of each flow is described as ID_F, which is a combination of the ID_{tot} of the respective information elements (according to the order illustrated in figure 1). Before data quality quantification, the database of an MFA study has to be inventoried and characterized according to the data characterization framework (Schwab et al. 2016). The procedure of quantitative estimation of data quality is described in the following. Exemplarily, the information defect of the flow F10 of the Pd MFA introduced above is quantified.

Data Attributes

Quality-relevant data attributes are listed in table 1 and exemplarily specified for the information elements F10.1 and F10.2 in the rightmost columns. Data attributes in text format (e.g., the producer type, which may be, among others, "national statistics" or "industrial association") have been translated to mathematically computable formats according to a translation

Table 1 Data attributes relevant for data quality evaluation selected from the article of Schwab and colleagues (2016) and their attribute numbers as identifiers

Data attribute	*Attribute no.*	*Designator*	*Scale*	*F10.1*	*F10.2*
No. of samples	*a104*	n	Absolute	1	1
Semantic precision	*a202*	a	Ordinal	0.3	0.2
Variety	*a211*	b	Ordinal	0.8	0.5
Disparity	*a212*	c	Ordinal	0.4	0.7
Producer type	*a308*	d	Ordinal	1	0.3
Origination type	*a311*	e	Ordinal	0.4	0.4
Origination quality	*a312*	f	Ordinal	0.7	0.5
Temporal variability	*a406*	g	Ordinal	0.2	0.2
Spatial variability	*a408*	h	Ordinal	1	0.1
Variability by further relation	*a410*	i	Ordinal	0	0
Temporal divergence	*a411*	j	Ordinal	1	0.1
Spatial divergence	*a412*	k	Ordinal	0.1	0.2
Further divergence	*a413*	l	Ordinal	0.2	0.0
Adaptation type	*a414*	m	Binary	0	1
Adaptation quality	*a415*	o	Ordinal	0.3	0

Note: The designators are used in the proposed formal procedure. Attributes are exemplarily specified for information elements F10.1 and F10.2 in the rightmost columns according to the framework proposed in the article of Schwab and colleagues (2016).

scheme provided in appendix S2-1 in supporting information S2 on the Web. Consequently, all attributes in table 1 are specified either on an ordinal scale from 0 to 1, where 0 means "good" and 1 means "bad," on an absolute scale (0, 1, 2 . . .), or on a binary scale (0 means yes, 1 means no).

In the first step of the evaluation procedure, the four information defects ID_i are quantified based on the 15 attributes listed in table 1.

Information Defects of Data Elements (ID_i)

The four information defects ID_i (ID_S, ID_R, ID_P, ID_C) are described as functions of data attributes (table 1). The information defect functions have been developed in a two-step heuristic procedure. First, the basic function type was chosen. Second, the relationship between data attributes, as qualified in the section *Approach and Conceptualization*, was formalized as a combination of data attributes by use of the chosen function type. The designators $a, b, \ldots, o$ of the attributes (table 1) are used in the functions presented in the following.

ID_S is regarded as a linear function of the attribute "semantic precision" (a, see table 1, where $a = 0$ represents data with unambiguous and clear meaning and $a = 1$ represents data with ambiguous or vague meaning), which means that the information defect is high when the meaning of data is vague (equation 1).

$$ID_S = a \tag{1}$$

ID_S of information element F10.2 equals the data attribute "semantic precision," so that $ID_{S,F10.2} = 0.2$.

The representativeness information defect ID_R is formalized as an exponential function of the attributes "variety" (b), "disparity" (c), and "number of samples" (n). ID_R increases with increasing variety and disparity (i.e., with increasing complexity of the described entity). ID_R and the information gain per additional sample decrease with increasing numbers of samples (equation 2). This relates to the equation of the standard error of the mean, where the error (expressed as standard deviation) decreases with increasing sample size (see Clark-Carter 2014).

$$ID_R = (\sqrt{b} * \sqrt{c})^{n/(n+1)} * \frac{(\sqrt{b} * \sqrt{c})}{\sqrt{n}} \tag{2}$$

The information element F10.2 refers to a complex entity with high variety and disparity and a small number of samples (see table 1), so that $ID_{R,F10.2} = 0.46$.

The provenance information defect is formalized as a function of the information producer (attribute "producer type" (d), first term in equation 3) and the way the data were collected (attributes "origination type" (e) and "origination quality" (f), second term in equation 3). The exponents determine the slope and the curvature of the function, that is, their nonlinearity. An exponent >1 results in a convex curved function. This means that ID_P is high only if both the information producer and the data generation method are specified with high attribute values. This way, data of a "bad" data producer are not *per se* evaluated as "bad" (as would be the case in a concave function, i.e., with an exponent <1) as long as a "good" data generation method was applied. Here, the exponents of the first and the second term are defined identically, that is, the information producer and the collection method have the same relative weight (equation 3).

$$ID_P = \left(d^{1.5} + \left(\frac{(e + f)}{2} \right)^{1.5} \right) \Big/ 2 \tag{3}$$

F10.2 was collected from a reputable scientific report which provides expert estimations on substance concentrations. The provenance information defect is, based on the attributes listed in table 1, $ID_{P,F10.2} = 0.23$.

The context information defect is formalized as a product of two constitutive parts. First, this is the degree to which data fits the system studied. Second, this is the quality nonadequate data were adapted to the system, for example, by scaling. In equation (4a), y denotes the data adequateness (ordinal (0-1); see equation 4b), m designates if data were adapted to the context (binary, yes/no, respective 0/1), and o refers to the adaptation quality (ordinal, 0 to 1).

$$ID_C = y - (1 - m) * (1 - o) * y \quad (4a)$$

This means that nonadequate data (expressed as y, first term) cause a high information defect, which decreases if these nonadequate data were well adapted to the context (second term).

The variable y in equations (4a) and (4b) denotes the degree to which data does not fit the context in three dimensions: time, space, and further (such as technology). It is a function of the divergence of the data from the system context ("divergence in three dimensions," j, k, l) and the variability of the data ("variability in three dimensions," g, h, i).

$$y = \left(\sqrt{g} * \sqrt{j} + \sqrt{h} * \sqrt{k} + \sqrt{i} * \sqrt{l}\right)/3 \quad (4b)$$

F10.2 does temporally and spatially diverge from the system boundary ("divergence"), but is little variable over time and space ("variability") given that the composition of flat screens in Austria and Germany can be regarded as quite similar in 2 subsequent years. The data have not been adapted to the system boundary. Considering the data attributes listed in table 1, it is $ID_{C,F10.2} = 0.09$. The information defect functions are visualized as surface plots provided in appendix S2-2 in supporting information S2 on the Web.

Information Defects of Information Elements (ID_{tot})

The four information defects are aggregated to a total information defect per information element (ID_{tot}) as Euclidian distance (the shortest connection of any point to the origin in an n-dimensional space) in a four-dimensional space. This is normalized to the measurement scale (0 to 1) by the number of information defects $ID_i = 4$ (equation 5).

$$ID_{tot} = \sqrt{\frac{\left(ID_S^2 + ID_R^2 + ID_P^2 + ID_C^2\right)}{4}} \quad (5)$$

In equation (5), ID_i are weighted by themselves. That means that an information element with a high defect in one dimension cannot be of overall good quality, even if the other three defects are low. When applied to F10.2 ($ID_{S,F10.2} = 0.20$, $ID_{R,F10.2} = 0.46$, $ID_{P,F10.2} = 0.23$, and $ID_{C,F10.2} = 0.09$), this results in $ID_{tot,F10.2} = 0.28$. The procedure presented thus far can be repeated for information element F10.1, so that $ID_{tot,F10.1} = 0.42$. The defect of information element F10.1 (number of flat screens sold) is higher than the defect of information element F10.2 (Pd content in flat screens). The information defect of flow F10 can now be expressed as one flow-specific value by combination of the two total information defects, $ID_{tot,F10.1}$ and $ID_{tot,F10.2}$.

Information Defects of Flows (ID_F)

ID_F is formalized as the square root of the sum of squares of all ID_{tot}, analogous to the combination of uncertainties in the Gaussian rule of error propagation (see the exponent in the denominator in equation 6, where z designates the number of information elements). This term can potentially increase indefinitely for increasing z and must consequentially be normalized to the measurement scale (0 to 1). Realistically, z is virtually never higher than four (a substance flow is typically quantified by multiplication of two information elements, quantity of goods times concentration, and, in fewer cases, by multiplication with additional information, such as, e.g., on volumes or areas). The term could be normalized by the square root of the numbers of information elements ($\sqrt{z}$). This straight forward normalization is not sensitive to the number of information elements z, and it averages the information defect of multiple information elements. However, it appears to be more suitable to consider that the more imprecise information there is to be combined, the less credible is the result. Having that in mind, ID_F can be also normalized by applying a logistic function such as the one proposed in equation (6). This function accumulates the information defects of multiple ID_{tot} per flow. A graphical comparison between normalization by $\sqrt{z}$ and a logistic function is provided in appendix S2-3 in supporting information S2 on the Web.

$$ID_F = \frac{1.5}{\left(1 + 2e^{-3\sqrt{\sum_{i=1}^{z} ID_{tot,i}^2}}\right)} - 0.5 \quad (6)$$

Applied to flow F10 with $ID_{tot,F10.1} = 0.42$ and $ID_{tot,F10.2} = 0.28$, this is $ID_{F10} = 0.54$.

Information Elements Specified by More Than One Data Element

The evaluation procedure has been illustrated for the situation of one data element per information element. As illustrated in figure 1, information elements may also be quantified based on more than one data element. For example, the information element F1.2 of the Pd study (Pd content of cars imported to Austria) consists of three data elements (see table 2 and supporting information S1 on the Web). That is, reference A states that the Pd content is A%, reference B says B%, and reference C says C%. Apparently, agents frequently introduce the mean of available data elements in their model when they cannot discriminate between the reliability of the three references.

In case of more than one data element per entity, ID_P and ID_C are calculated on the level of data elements and ID_S and ID_R are calculated on the level of information elements (i.e., per entity). That is, because each data element may have a different provenance (different ID_P) and may be of different adequateness to the context (different ID_C), but is used to represent the same entity (same ID_R) with the same meaning (same ID_S). This becomes clear when reconsidering the concept presented in figure 2.

Table 2 Information element F1.2 of the Pd study consists of three data elements

Information element	*Data element*	ID_S	ID_R	ID_P	ID_C
F1.2		0.20	0.23		
	F1.2a			0.25	0.05
	F1.2b			0.28	0.22
	F1.2c			0.23	0.18
$ID_{i,F1.2}$		0.20	0.23	0.23	0.05

Note: The lowest ID_P and ID_C are selected for further processing in equation (5).
Pd = palladium.

Equation 5 requires a set of four information defects ID_i per information element. To formulate such a set of four ID_i for the situation presented in table 2, a straightforward approach is chosen. Experience shows that, typically, an agent only introduces additional data elements per information element when expecting information gain (e.g., by taking the mean of two independent references). This is considered here, and the lowest ID_P and ID_C are selected from the set of provenance and context defects (see table 2, where F1.2 is quantified based on $n = 3$ data elements). Consequently, the information defect $ID_{tot,F1.2}$ decreases (because of min ID_P and ID_C and also because $n>1$ in ID_R), which reflects the information gain.

The information defect approach results in a new quantity for evaluation of regional MFAs. This new quantity indicates the reliability of model input data and enables distinguishing material flows by their data quality. The evaluation procedure is applied to all flows of the Pd MFA in the following.

Case Study on Palladium Flows in Austria 2011

The presented procedure for quantitative estimation of data quality is applied to the Pd flow system illustrated in figure 3. For a more detailed description of the Pd MFA and a quantitative diagram, please refer to the article of Laner and colleagues (2015a).

Information defects ID_i per data element, ID_{tot} per information element, and ID_F per flow are computed according to equations (1) to (6). The results are illustrated in figure 4. A detailed table of all information defects is provided in appendix S2-4 in supporting information S2 on the Web.

With regard to the concept of information defects, low bars indicate good data quality (flows F1 to F3) and high bars indicate poor data quality (flows F4 to F14, F19, and F25). Information defects higher than 0.5 signify data of considerably poor quality. For some flows, no input data were available. Clearly, nonexistent information cannot be defective, but complete ignorance can be regarded as a maximum information defect. Thus, $ID_F = 1$ is assigned for unknown flows (flows F15 to F18 and F20 to F24). The bars in figure 4 denote the *a priori* knowledge about flows, that is, the knowledge before application of a material flow model. By balancing of flows in a model (in the Pd study, the STAN algorithm [www.stan2web.net] was applied), initially unknown flows are calculated and the *a posteriori* state of information differs from the *a priori* information state. Thus far, the information defects enable assessing the state of information about a material flow system and underpin qualitative observations about available information by quantitative means (subsequent applications of information defects in material flow modeling are outlined in the *Concluding Remarks and Outlook* section). For example, data quality is often found to decrease over the life cycle of materials and to be better for sectors of economic interest, such as trade and manufacturing, in contrast to the consumption and waste management sectors (see, among others, Mao et al. 2008; Graedel et al. 2004). The results of the Pd case study indicate that data quality is considerably better at the system input side, whereas data quality of flows to the environment (such as dissipative fate; flow 25) is poor. For many flows in the waste management sector (e.g., flows 20 and 21), no data are available. The information defects do now provide an opportunity to quantitatively express data quality and illustrate weaknesses and tendencies of the database in a systematic and reproducible way.

Discussion of the Formalization Procedure

Data attributes can be mathematically combined in many different ways for specification of information defects. The formalizations of ID_i, ID_{tot}, and ID_F proposed here deliver mathematically sound and reasonable results for quantitative data quality evaluation. They have been selected from a number of possible formalizations based on comprehensive qualitative and quantitative tests, where individual steps of the quantification procedure have been varied and compared regarding their absolute output and their relative ranking based on Monte Carlo simulation, surface plots, and correlation analysis (Schwab and Rechberger 2015). The mathematically simplest alternative approach is to formalize the defect of information elements as an average (denoted as $ID_{tot,average}$ in the following) of all ordinal attributes (table 1). In figure 5, this $ID_{tot,average}$ is plotted against ID_{tot} of the information elements of the Pd case study.

The averaged information defect appears to equalize the results and deviate from ID_{tot}, especially for increasing information defects. Although $ID_{tot,average}$ is mathematically feasible, it is of little meaning with regard to the information defects. That is, because some data attributes are obviously related to others, which is not considered by $ID_{tot,average}$. For example, the attribute "temporal divergence" interacts with the attribute "temporal variability" when it comes to data quality evaluation as it is obvious that outdated data (temporal divergence) is only defective if the data vary over time (temporal variability). The example of $ID_{tot,average}$ indicates that the adequateness of very simple mathematical formalizations to express information defects may be limited.

To assess and discuss the results of the data quality approach, the information defects of the Pd case study are also compared

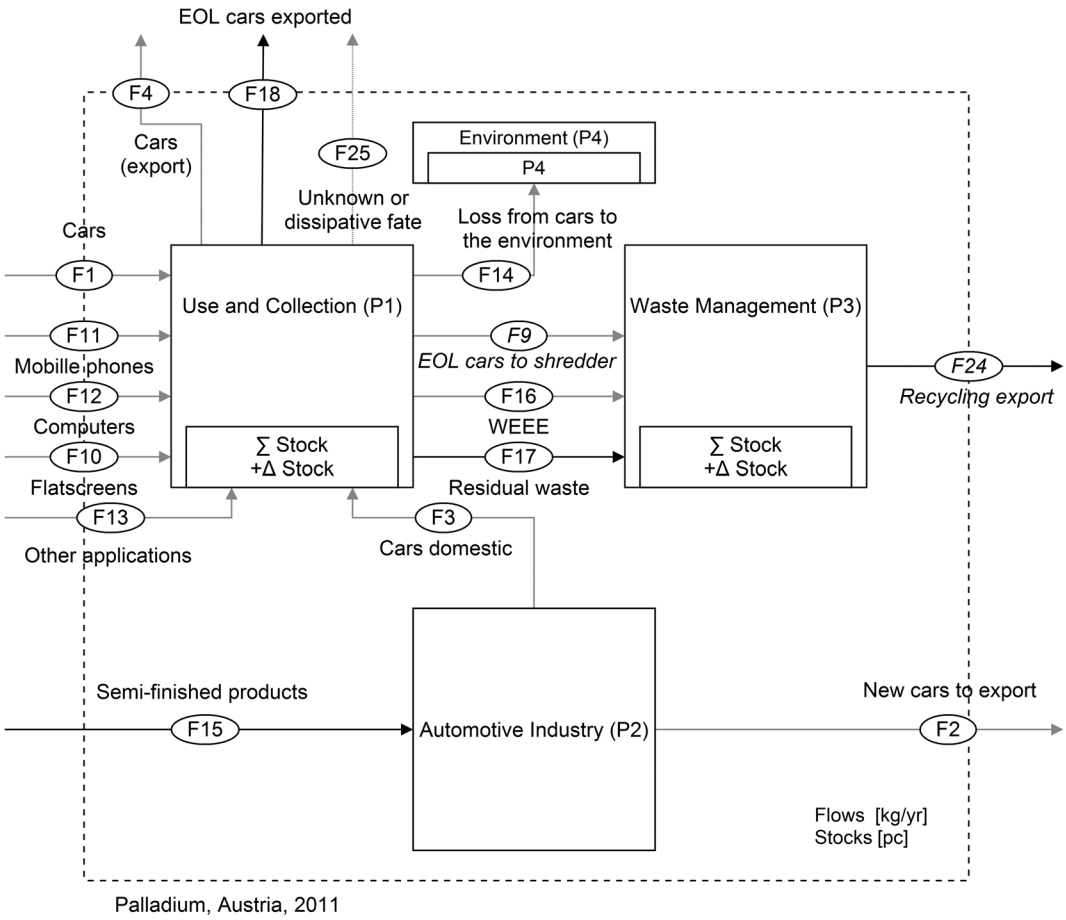


Figure 3 Structure of the 2011 Austrian Pd MFA (Laner et al. 2015a). The system consists of 25 flows (16 in the main system and nine in subsystems "use and collection" and "waste management"). EOL = end of life; kg/yr = kilograms per year; MFA = material flow analysis; Pd = palladium.

to data uncertainties, as calculated in Laner and colleagues (2015a) (see figure 5). Laner and colleagues used an adapted version of Hedbrant and Sörme (2001) for calculation of data uncertainties in the Pd study. Whereas this is based on categorization of data into five quality categories according to their origination, the information defect approach distinguishes data quality by a higher number of data characteristics and considers interconnections between data attributes. Figure 5 indicates that uncertainty calculations and ID_{tot} can differ, but show a similar trend (Spearman rank correlation coefficient between uncertainty ranges and ID_{tot} in the comparison presented is $\rho = .7$; between ID_{tot} and $ID_{tot,average}$ it is $\rho = .8$). The range covered by ID_{tot} seems less wide than the range covered by the uncertainty estimates (figure 5).

A difference between the introduced method and other approaches to data quality, such as the one in Laner and colleagues (2015b), is that data quality is not formalized based on static indicators and categories. Data quality is here formalized in a model-type setup, where different data characteristics are linked and may enhance or reduce the resulting information defect, depending also on the magnitude of related attributes. The data attributes contribute to the information defects ID_i, ID_{tot}, and ID_F to a variable extent, depending on the model formalization. The weight of data attributes in ID_F in the formalization proposed above has been analyzed. This was done by investigating the relative impact of variations in inputs (individual data attributes; table 1) on the observed variation of the output (ID_F; equation 6) in a sensitivity analysis (Monte Carlo–based multiple linear regression). The relative weight of data attributes is displayed in figure 6.

According to the formalization presented here, the precise knowledge of the meaning of data (ID_S) and the provenance of data (ID_P) contribute most to ID_F. In contrast, all input attributes considered would have the same weight in the alternative formalization $ID_{tot,average}$ mentioned earlier. Certainly, however, the weight of data attributes in the information defects can be varied, for example, by the introduction of weighting factors in equations (5) and (6), or by modification of equations (1) to (4).

The application of the data quality evaluation procedure may require more time than other approaches (see the overview of existing methods in Laner et al. [2015b]). In return, it enables better understanding of the factors determining the information quality of material flows. For convenient and time-saving

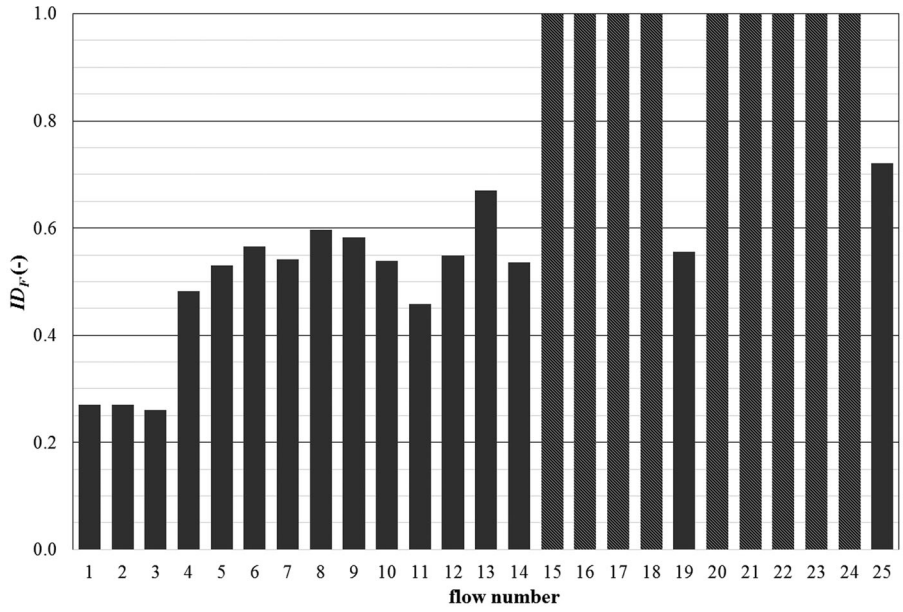


Figure 4 Data quality of the flows in the Pd MFA expressed as information defects ID_F. Low defects indicate good data quality; high defects indicate poor data quality. Flows without input data are here assigned $ID_F = 1$. MFA = material flow analysis; Pd = palladium.

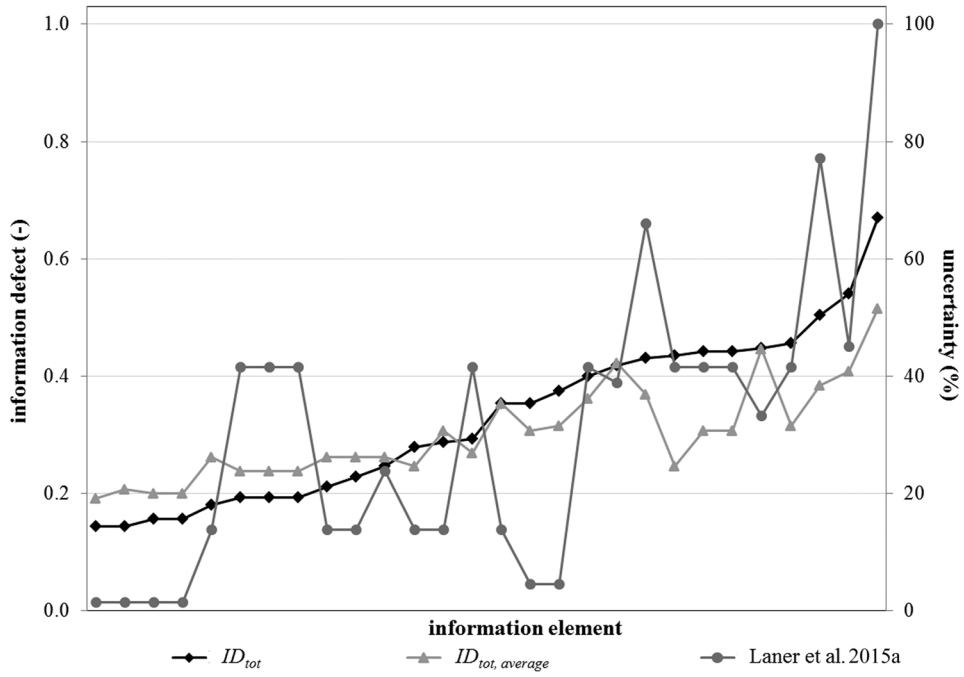


Figure 5 Comparison of ID_{tot} to an alternative total information defect $ID_{tot,average}$ and data uncertainty estimations of Laner and colleagues. Information defects (dimensionless) are plotted on the primary, uncertainties (%) on the secondary y-axis. The values are sorted according to increasing ID_{tot}. The connecting lines between the points are introduced to enable simple comparison of the plotted options.

application, the model presented here is implemented in a spreadsheet tool attached to the data characterization matrix (Schwab et al. 2016), which is provided in supporting information S1 on the Web. In that spreadsheet tool, data quality can be calculated automatically once a database has been characterized. Evaluation of a study with the extent of the Pd case presented here requires 30 to 40 work hours. More detailed full-scale national resource budgets (the Pd study was not full scale; it had a focus on end of life of consumer product flows) may require more time for their characterization and evaluation (a possible general system structure for these kind of regional MFAs is proposed, e.g., in Pauliuk et al. [2015]). It may be difficult to retrieve all information necessary for sound data characterizations of existing studies that used other data

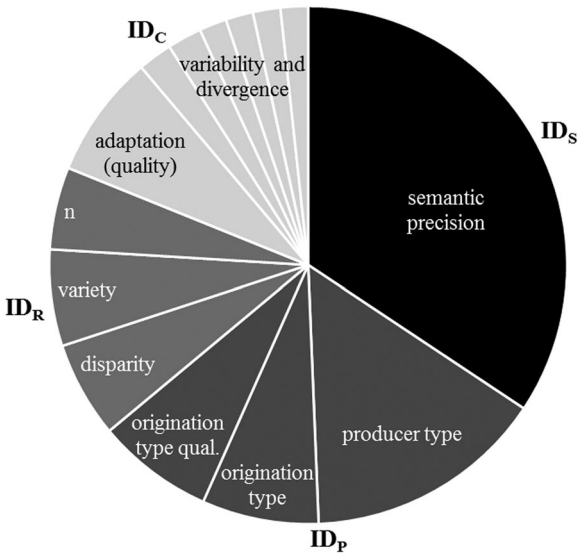


Figure 6 Relative weight of individual data attributes and ID_i in information defect ID_F.

documentation schemes or that provide no complete and consistent data documentation. Thus, it is beneficial to apply the data characterization framework while preparing a study, that is, in parallel to data collection.

Concluding Remarks and Outlook

Objective possibilities for data quality evaluation are usually limited in MFA. Data quality evaluation is inevitably subjective, also with the proposed procedure. Nevertheless, it reduces the influence of author opinions and subjectivity by systemizing the evaluation procedure and by moving choices from generic data classifications to an evaluation of individual data attributes. Despite the systematic approach, agents with different backgrounds and with different degrees of knowledge may characterize data and their attributes differently and may thus produce differing results in data quality evaluation, especially when the approach is applied not in parallel, but subsequent to a study. The method presented should be seen as a transparent "best guess" procedure for facilitating reproducible and transparent expert estimations on the abstract phenomenon "data quality."

The output of the method presented is a ranking of flows on an ordinal scale according to their data quality. An alternative concept of Laner and colleagues (2015b) aims at providing uncertainty ranges for MFA data. Often, the initial idea behind uncertainty ranges is to express the reliability of the underlying data. Naturally, it may be difficult to express epistemic uncertainties ("lack of knowledge") by absolute measures and estimations on the dimension and geometry of uncertainty ranges may be highly speculative. That is why the approach presented here is designed to evaluate the degree of credibility of *a priori* data without simulating an absolute quantification of uncertainty. Information defects can, in subsequent work, be applied as dimensionless factors in characterization functions of material flow models and as indicators for the reliability of data (e.g., in Laner et al. 2015a) or as factors in data reconciliation algorithms (e.g., Kopec et al. 2015). The information defects can be applied as indicators for epistemic uncertainty in data uncertainty frameworks (such as, among others, Dubois and Guyonnet [2011] and Clavreul et al. [2013]).

Uncertainty ranges are practicable and often desired by material flow analysts. The dimensionless information defects can also be translated to uncertainty ranges by application of scaling functions and by multiplication with a coefficient of variation or an uncertainty factor. More than that, it is possible to test whether the integration of empirically derived probabilities into the information defect concept is adequate for particular MFA applications. Given that statistical characteristics such as dispersion measures or probability distributions are, if available, also part of the data characterization matrix (see attribute group "statistical characteristics" in the data characterization matrix of Schwab et al. [2016]), this could also be exploited for characterization of observed variability. In contrast to the information defects, which indicate the epistemic uncertainty, natural variability is not a component of uncertainty. Rather, variability is an intrinsic property of any entity with more than one realization and thus also an intrinsic part of a complete piece of information (e.g., as an empirically derived distribution). Not to know about variability, in return, is a knowledge shortcoming and thus epistemic uncertainty. As such, it is also part of the approach presented here (see the representativeness information defect (ID_R) in equation (2), which increases with decreasing number of samples n).

In combination with the data characterization framework presented in a previous article (Schwab et al. 2016), the evaluation procedure proposed enables the documentation, characterization, evaluation, and communication of the information basis of regional MFAs. The information defects indicate the reliability of data and help to find weak points in the data structure. They enable identifying the reason for data weaknesses (Is the source unreliable? Is the number of samples not high enough? Is the meaning of the data unclear?) and aid in adopting adequate measures for filling data gaps. When not interpreting information defects as factors for uncertainty evaluation, but leaving them as dimensionless measures for a "state of knowledge," the results can be applied for comparing regional MFAs of different substances, regions, or years to one another and for measuring the learning effect on regional material flow systems over time.

Acknowledgments

This research was funded by the Austrian Federal Ministry of Science, Research and Economy. The authors thank Nađa Džubur and Oliver Cencic, Technische Universität Wien, for their contributions to this study, and three anonymous reviewers for useful comments to a previous version of this article.

References

Bonnin, M., C. Azzaro-Pantel, L. Pibouleau, S. Domenech, and J. Villeneuve. 2013. Development and validation of a dynamic material

flow analysis model for French copper cycle. *Chemical Engineering Research and Design* 91(8): 1390–1402.

Cencic, O. and R. Frühwirth. 2015. A general framework for data reconciliation—Part I: Linear constraints. *Computers & Chemical Engineering* 75: 196–208.

Clark-Carter, D. 2014. Standard error. In *Wiley StatsRef: Statistics reference online*. Chichester, UK: John Wiley & Sons Ltd.

Clavreul, J., D. Guyonnet, D. Tonini, and T. Christensen. 2013. Stochastic and epistemic uncertainty propagation in LCA. *The International Journal of Life Cycle Assessment* 18(7): 1393–1403.

Dubois, D. and H. Prade. 2010. Formal representations of uncertainty. In *Decision-making process: Concepts and methods*. New York: Wiley-ISTE.

Dubois, D. and D. Guyonnet. 2011. Risk-informed decision-making in the presence of epistemic uncertainty. *International Journal of General Systems* 40(2): 145–167.

Dubois, D., H. Fargier, M. Ababou, and D. Guyonnet. 2014. A fuzzy constraint-based approach to data reconciliation in material flow analysis. *International Journal of General Systems* 43(8): 787–809.

Egle, L., O. Zoboli, S. Thaler, H. Rechberger, and M. Zessner. 2014. The Austrian P budget as a basis for resource optimization. *Resources, Conservation and Recycling* 83: 152–162.

Gottschalk, F., R. W. Scholz, and B. Nowack. 2010. Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO_2 particles. *Environmental Modelling & Software* 25(3): 320–332.

Graedel, T. E., D. van Beers, M. Bertram, K. Fuse, R. B. Gordon, A. Gritsinin, A. Kapur, et al. 2004. Multilevel cycle of anthropogenic copper. *Environmental Science & Technology* 38(4): 1242–1252.

Hedbrant, J. and L. Sörme. 2001. Data vagueness and uncertainties in urban heavy-metal data collection. *Water, Air and Soil Pollution: Focus* 1(3–4): 43–53.

Klinglmair, M., O. Zoboli, D. Laner, H. Rechberger, T. Fruergaard Astrup, and C. Scheutz. 2016. The effect of data structure and model choices on MFA results: A comparison of phosphorus balances for Denmark and Austria. *Resources, Conservation and Recycling* 109(109): 166–175.

Kopec, G. M., J. M. Allwood, J. M. Cullen, and D. Ralph. 2015. A general nonlinear least squares data reconciliation and estimation method for material flow analysis. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12344.

Laner, D., H. Rechberger, and T. Astrup. 2014. Systematic evaluation of uncertainty in material flow analysis. *Journal of Industrial Ecology* 18(6): 859–870.

Laner, D., H. Rechberger, and T. Astrup. 2015a. Applying fuzzy and probabilistic uncertainty concepts to the material flow analysis of palladium in Austria. *Journal of Industrial Ecology* 19(6): 1055–1069.

Laner, D., J. Feketitsch, H. Rechberger, and J. Fellner. 2015b. A novel approach to characterize data uncertainty in MFA and its applications to plastic flows in Austria. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12326.

Mao, J. S., J. Dong, and T. E. Graedel. 2008. The multilevel cycle of anthropogenic lead: I. Methodology. *Resources, Conservation and Recycling* 52(8–9): 1058–1064.

Nakajima, K., H. Ohno, Y. Kondo, K. Matsubae, O. Takeda, T. Miki, S. Nakamura, and T. Nagasaka. 2013. Simultaneous material flow analysis of nickel, chromium, and molybdenum used in alloy steel by means of input-output analysis. *Environmental Science & Technology* 47(9): 4653–4660.

Patrício, J., Y. Kalmykova, L. Rosado, and V. Lisovskaja. 2015. Uncertainty in material flow analysis indicators at different spatial levels. *Journal of Industrial Ecology* 19(5): 837–852.

Pauliuk, S., G. Majeau-Bettez, and D. B. Müller. 2015. A general system structure and accounting framework for socioeconomic metabolism. *Journal of Industrial Ecology* 19(5): 728–741.

Rechberger, H., O. Cencic, and R. Frühwirth. 2014. Uncertainty in material flow analysis. *Journal of Industrial Ecology* 18(2): 159–160.

Schwab, O. and H. Rechberger. 2015. Ermittlung des Datenbedarfs für Nationale Rohstoffbilanzen— Entwicklung von SFA-spezifischen Datenqualitätsindikatoren [Investigation of the data requirements of national resource budgets—Development of MFA-specific data quality indicators]. Report for the Austrian Federal Ministry of Science, Research and Economy. Vienna: Technische Universität Wien.

Schwab, O., O. Zoboli, and H. Rechberger. 2016. A data characterization framework for material flow analysis. *Journal of Industrial Ecology* DOI: 10.1111/jiec.12399.

Wu, H., Z. Yuan, Y. Zhang, L. Gao, S. Liu, and Y. Geng. 2014. Data uncertainties in anthropogenic phosphorus flow analysis of lake watershed. *Journal of Cleaner Production* 69(15): 74–82.

About the Authors

Oliver Schwab is a research associate, **David Laner** is a senior researcher, and **Helmut Rechberger** is a professor for resource management at the Institute for Water Quality, Resource and Waste Management at Technische Universität Wien, Vienna, Austria.

Supporting Information

Supporting information is linked to this article on the *JIE* website:

Supporting Information S1: The supporting information S1 file (.xlsx) provides the following: 1) a characterized information inventory of the 2011 Austrian Palladium budget characterized in the data characterization matrix (DCM) and 2) sets of information defects for all flows of the Palladium study, quantified based on the attributes of the DCM according to the procedure described in the article.

Supporting Information S2: The supporting information S2 document provides the following: 1) a translation scheme for text format data attributes to computable scales, 2) surface plots of information defect functions, 3) a graphical comparison of two ID_F normalization functions, and 4) a full list of the case studies' information defects.

Information Content, Complexity and Uncertainty in Material Flow Analysis

Oliver Schwab* and Helmut Rechberger

Institute for Water Quality, Resource and Waste Management, Technische Universität Wien, Karlsplatz 13/226, 1040 Vienna, Austria

Revised manuscript as submitted to Journal of Industrial Ecology

Abstract

Material Flow Analysis (MFA) is a useful method for modeling, understanding and optimizing metabolic systems. Among others, MFAs can be distinguished by two general system properties: First, they differ in their complexity, which depends on the system structure and the system size. Second, they differ in their inherent uncertainty, which arises from limited data quality. In this article, uncertainty and complexity in MFA are approached from a system-theoretical perspective and expressed as formally linked phenomena. MFAs are, in a graph-theoretical sense, understood as networks. The uncertainty and complexity of these networks are computed by use of information measures from the field of theoretical ecology. The size of a system is formalized as a function of its number of flows. It defines the potential information content of an MFA system and holds as a reference against which complexity and uncertainty are gauged. Integrating data quality measures, the uncertainty of a quantitative MFA before and after balancing is determined. The actual information content of an MFA is measured by relating its uncertainty to its potential information content. The complexity of a system is expressed based on the configuration of each individual flow in relation to its neighboring flows. The proposed metrics enable different systems to be compared to one another and the role of individual flows within a system to be assessed. They provide information useful for the design of MFAs and for the communication of MFA results. For exemplification, the regional MFAs of aluminum and plastics in Austria are analyzed.

Keywords: Material Flow Analysis, network, information content, uncertainty, complexity, Industrial Ecology

<heading level 1> Introduction

Material flow analysis (MFA) provides useful information on metabolic systems that span natural, technological and economic environments. Procedures for preparation of MFAs and tools for their representation have been widely harmonized. MFAs of various scopes and materials have proven useful in scientific discourse and for decision making in industrial and institutional contexts (see, for example, Morf and Brunner (1998), Velis et al. (2013), Trinkel et al. (2015), Zoboli et al. (2016)). MFA incorporates databases of increasing size and quality and reveals more and more details on socio-metabolic systems. In that course, it has been argued that, in Industrial Ecology, uncertainty and complexity issues are increasingly relevant (Kay 2002) and that information is a notable phenomenon, also because "uncertainties paralyze us" and because it is not clear how much information is needed for design of systems (Bettencourt and Brelsford 2015). This holds also for MFA: When recalling that studies of material flow systems both reveal new information and depend on existing information when being prepared (Chen and Graedel 2012), the dual role of information in MFA becomes apparent. For making informed MFA-based decisions, agents are not only to know about MFA results on a material level, but also about their reliability, that is, about the „uncertainty" and, respectively, the „information content" of a given material flow system. For making good use of available data, agents also are to know about the "complexity" of a material flow system, as increasingly complex systems require, in comparison to systems of more trivial structures, increasing amounts of information in order to be solved. In this article, the phenomena uncertainty and complexity are addressed as properties of descriptive MFAs (MFAs which aim on understanding temporally and spatially precisely defined systems such as the flows of a material *X* in region *Y* in the year *Z*). The information content of material flow systems is derived from their uncertainty.

<heading level 2> Uncertainty in Material Flow Analysis

In Industrial Ecology, established statistical procedures such as stochastic modeling and scenario modeling are often applied for the treatment of uncertainties, for example in input-output models (Lenzen et al. 2010), in Life Cycle Assessment (Lloyd and Ries 2007) and in MFA (Gottschalk et al. 2010). Often, these approaches require more information than is actually available as data are typically given in the form of individual, isolated values and not in the form of statistically exploitable datasets. This holds especially for MFAs such as the two cases presented later in this article, where data uncertainty relates to knowledge shortcomings ("epistemic uncertainty", Laner et al. (2014)). Consequently, even though

methods for treatment of known data uncertainties in MFA are available (see, for example, Kopec et al. (2015) and Cencic (2016)), means for actual characterization and representation of data uncertainty in the absence of statistical evidence are limited.

As alternatives to author judgements or expert estimates of uncertainties (as, for example, performed in Graedel et al. (2004), Huang et al. (2007) and Ott and Rechberger (2012)), more systematic and transparent approaches have been proposed. In a concept of Hedbrant and Sörme (2001), MFA data are assigned to five uncertainty levels according to their origin, and this classification is then translated to uncertainty ranges. Expanding on that idea and integrating elements of the LCA-specific data quality concept of Weidema and Wesnæs (1996), Laner et al. (2015b) propose a concept in which data uncertainties are formalized as functions of five data quality indicators. In an approach of Schwab et al. (2016a), data quality is expressed by means of multidimensional functions of data characteristics as belief indicators named "information defects" (*ID*). These *ID*s reflect the degree of belief in given information to be true in a particular MFA context. All these approaches are geared towards systematic estimations for *a priori* characterization of data when statistical information is absent. A difference is, that the approaches of Hedbrant and Sörme (2001) and of Laner et al. (2015b) aim on quantification of uncertainty ranges, while the approach of Schwab et al. (2016a) aims on dimensionless data quality indicators, where low information defects ($ID \rightarrow 0$) relate to data of good quality, high information defects ($ID \rightarrow 1$) relate to data of poor quality and $ID=1$ relates to complete ignorance.

Material flow analysts often choose to represent data uncertainty by means of uncertainty ranges, also because these can be treated in established frameworks. If no information on statistical variability is provided, however, the idea behind uncertainty ranges is to express the degree of belief an agent has in given data to be true, although it may be difficult to specify uncertainty ranges in the absence of empirical evidence. As a consequence, besides the choice of distribution geometries, the specification of uncertainty ranges is probably arbitrary: Why is an uncertainty range of ± 20% for quantity *X* assumed and not a range of ± 30%? Is ± 100% a natural upper limit, or is that often rather chosen because of mathematical convenience and the physical constraint that the lower bound of a quantity is zero? The question remains whether it is useful to quantify the unquantifiable, that is, to provide quantitative uncertainty ranges when these are actually unknown, also because this conveys the unjustified impression of empirical evidence. With the incentive to avoid the use of uncertainty ranges but to still allow for both relative and absolute comparisons, in this article, uncertainty is regarded as a

system property of MFAs which involve imperfect information. As formalized later in this study, the potential uncertainty of a system increases with its number of flows and decreases when more and better data is incorporated.

<heading level 2> Complexity in Material Flow Analysis

Complexity concepts are of increasing interest for systems analysis in Industrial Ecology, as put together in two special issues of this journal (Dijkema and Basson (2009), Dijkema et al. (2015); see the respective review articles Wood and Lenzen (2009) and Meerow and Newell (2015)). Although the term "network" is frequently used in a qualitative sense (Heijungs 2015), graph theory is a rich source of concepts for quantitative analysis of network structures and parallels to analytical approaches in economics and in ecology have been revealed (Suh 2005). Many graph-theoretical applications draw from theoretical ecology (Odum 1994; Ulanowicz 1997) and analogies between ecosystems and social, economic and industrial systems have been identified (Côté and Hall 1995; Graedel 1996; Korhonen 2001; Bailey et al. 2004). As reviewed in Schiller et al. (2014), graph-theoretical network measures have been applied for describing system structures in Industrial Ecology and for comparing different systems to one another (for a recent application, see Nuss et al. (2016)). Despite the increasing use for analysis of non-trivial structures, graph-theoretical network measures such as "connectedness", "clustering" or "cyclicity" have in few studies been specifically interpreted as relative complexity measures, for example regarding life cycle inventories (Navarrete-Gutiérrez et al. 2015) or industrial ecosystems (Layton et al. 2016). In MFA, complexity measures have not been specifically addressed so far, although graph-theoretical complexity measures are also applicable in MFA. When regarding complexity in the sense of "static complexity", which refers to the "number of parts and their linkages" (Allenby 2009), it appears to be useful to express complexity not as a merely relative, but as an absolute measure, as systems (also material flow systems) may not only differ in their complexity because of different linkage patterns, but also because of varying system sizes. An alternative for analysis of network complexity is presented in this article. It is elaborated specifically for MFA systems and is, as other approaches to complexity in Industrial Ecology, inspired by theoretical ecology.

As material flow systems today cover increasing numbers of materials and regions, there is an interest in identifying similarities and differences of MFA systems (Klinglmair et al. 2016). It has been observed in different fields of Industrial Ecology that differences in system structure are to a varying degree to be attributed to actual differences in physical systems, but also to

priorities of modelers, the chosen level of detail and the structure of available data (Heijungs 2015). Until now, both comparisons of MFA system structures and evaluations of the impact of MFA input data on MFA results are often limited to qualitative considerations (Klinglmair et al. 2016). A set of measures for specifying and comparing MFA systems by quantitative rather than sheer qualitative means appears to be helpful to facilitate further comparisons of MFA systems and also for communication of MFA results. In this work, such measures are proposed. With a focus on flows, the uncertainty and complexity of MFA systems are formalized as properties of MFA systems and quantitatively expressed in the same, abstract dimension. The information content of MFA systems is derived from their uncertainty. The formal framework used for computation is borrowed from the field of theoretical ecology, as introduced in the following.

<heading level 1> Information measures in theoretical ecology

In theoretical ecology, a concept for the description of networks has been elaborated based on work of Rutledge et al. (1976). The concept is constructed around the narrative that the functioning of ecosystems can be understood by means of information theory as a function of system size, proportions of flows in relation to other flows and system structure. Based on these ideas, Ulanowicz (1980) developed a set of aggregate measures for describing the state of ecosystems and their potential to undergo change. The starting point is a perspective on an ensemble of flows, which is, in the notation of Ulanowicz et al. (2009), expressed as

$$H = -k \sum_{ij} \frac{T_{ij}}{T_{..}} \, log \, \frac{T_{ij}}{T_{..}} \qquad \text{Eq. 1}$$

where T_{ij} refers to a flow from agent i to j, $T_{..}$ (a dot refers to summation over an index) refers to the aggregate of all flows in the system and k refers to a positive scaling constant. The measure is used to refer to "system diversity" (Rutledge et al. 1976) or "system capacity" (Ulanowicz 1997). Rutledge et al. (1976) argue that H can be decomposed into two components based on information on each flows' configuration in the system, so that, in the notation brought forward by Ulanowicz et al. (2009),

$$X = k \sum_{ij} \frac{T_{ij}}{T_{..}} \, log \, \frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \qquad \text{Eq. 2}$$

and

$$\psi = -k \sum_{ij} \frac{T_{ij}}{T_{..}} \, log \, \frac{T_{ij}^2}{T_{i.} T_{.j}} \qquad \text{Eq. 3}$$

where $T_{i.}$ refers to the aggregate quantity that leaves *i*, $T_{.j}$ refers to the aggregate quantity that enters *j* and T_{ij} refers to the quantity that both leaves *i* and enters *j*. *X* and ψ relate to the information-theoretical concepts of mutual information and conditional entropy, which are measures of association for quantifying the relationship between two variables. Eq. 2 and Eq. 3 express the association of agents *i* and *j* as a function of the flow of matter from *i* to *j* in relation to the aggregate quantities leaving *i* and entering *j*. High *X* specifies that *j* depends mainly on *i*, high ψ specifies that *j* depends little on *i* but mainly on other adjacent *i*'s. The measures have been applied to ecosystems, for example for examining the functioning of food webs (Wulff et al. 1989; Baird and Ulanowicz 1989) or for comparison of ecosystems (Christian et al. 2005). It has also gained interest in other system-oriented fields such as economics (Goerner et al. 2009) and Industrial Ecology (Kharrazi et al. 2013), where it has been interpreted as a measure for the sustainability of a network. As highlighted by Kharrazi et al. (2013), it is a useful characteristic of the measures that they allow considering both intensive and extensive system properties, a feature which is utilized later in this work.

The measures for the uncertainty and complexity of MFAs of a given size and structure proposed in this article have a focus on the role of flows. The measures are based on variables relating to the quality of flow data and to the configuration of flows in the system. Their computation is facilitated by Eq. 1, Eq. 2 and Eq. 3. In order to provide a quantity that holds as a reference against which the uncertainty and complexity of a given MFA system can be measured, Eq. 1 is reformulated to express the "informational" system size *S* of an MFA as a function of the number of flows n_F in a system where, for now, all flows *Fi* have the weight 1, and the aggregate of all flows in the system $\sum_i Fi$ (here: $\sum_i Fi = n_F$) is used as the scaling constant *k*, so that

$$S = -\sum_i Fi \cdot \sum_{(Fi)} \frac{Fi}{\sum_i Fi} \log \frac{Fi}{\sum_i Fi} = -n_F \, log \frac{1}{n_F} \qquad \text{Eq. 4}$$

S is a monotonic increasing function of n_F and for systems with an arbitrary number of flows, it is $\lim_{n_F \to \infty} S = \infty$. A binary logarithm is chosen for computation in this article and the resulting quantity is referred to as "informational units". Each individual flow contributes to the magnitude of *S*, i.e., is a component of the sum, which allows quantitatively specifying the contribution of any individual *Fi* to the aggregate system uncertainty and system complexity, as elaborated in the following.

<heading level 1> Uncertainty of material flow systems

A typical procedure for filling a given qualitative MFA system with numbers consists of two steps (Brunner and Rechberger 2016). In the first step, *a priori* data for specification of system variables is collected. This data may be incomplete and inconsistent and therefore, in the second step, is balanced and reconciled in an MFA model. This second step increases the completeness and decreases the inconsistencies of data in the model. Ideally, such balanced MFAs provide reliable information on material flow systems. If all flows in a system were known with absolute certainty, their information content would be maximal. It would increase with the level of detail of a given system, that is, with the number of flows that are distinguished and correctly specified. As such ideal cases are unrealistic because of data quality limitations, there typically is a remaining degree of uncertainty in MFA results (Laner et al. 2014). As statistical evidence for specification of data quality is frequently limited in MFA, it can be expressed by reliability indicators such as the information defects mentioned earlier in this article. Because of data quality limitations and the resulting uncertainty in the system, the actual information content of given MFA systems usually is lower than their potential information content. A formal way to express this limitation by quantitative means is proposed in this section.

<heading level 2> Uncertainty of systems with *a priori* data

As information can be understood as the absence of uncertainty, and *vice versa*, *S* allows two interpretations. First, if all flows were known, it can be interpreted as the potential information content of a material flow system. Second, it can be interpreted as the uncertainty of a given qualitative material flow system, where none of the flows is known. This uncertainty may be reduced by integrating data on these unknown flows into the system. In other words, the uncertainty of a system with *a priori* data on flows (U_{ap}) may be expressed as a composite of *S*, which reflects the uncertainty of a system without data. Instead of assigning the equal weight 1 to all flows, as it is in Eq. 4, the flows can be weighted by their information defects ID_{Fi}. The uncertainty of a system with *a priori* information (U_{ap}) can then be formulated as

$$U_{ap} = -\sum_{i} Fi \cdot \sum_{i=1}^{n_F} \frac{ID_{Fi}}{\sum_i Fi} \; log \frac{ID_{Fi}}{\sum_i Fi} = -\sum_{i=1}^{n_F} ID_{Fi} \; log \frac{ID_{Fi}}{\sum_i Fi} \qquad \text{Eq. 5}$$

The measure U_{ap} refers to the uncertainty remaining in a quantitative MFA system after data of varying quality on flows is considered. It is $\lim_{ID \to 0} U_{ap} = 0$ and $U_{ap} \leq S$. This becomes clear when considering the simple examples in Table 1 and the case studies later in this article. The observation that the uncertainty of flows may better be expressed in relation to the uncertainty of other flows in a system (Klinglmair et al. 2016) reflects in Eq. 5, where the uncertainty of each individual flow is not only a function of its specific *ID*, but also expressed in relation to the sum of all *ID* in the system.

<heading level 2> Uncertainty of balanced material flow systems

By balancing material flow systems, conflicting model input data is reconciled and data gaps are closed. As a result, the uncertainty of a system decreases (i.e., the consistency of a system increases) when it is balanced. Consequently, the uncertainty after balancing (U_b) should be lower than the uncertainty before balancing (U_{ap}). A typical application for balancing material flow systems is the software STAN (www.stan2web.net). In STAN, an algorithm based on the weighted least square method is implemented for data reconciliation. A system with information on flow quantities (linear constraints) is reconciled based on the relation between factors such as standard errors (see Cencic (2016)). In this article, the information defects are used as factors in data reconciliation with software STAN.

The uncertainty remaining in a system after balancing (U_b) is computed by replacing ID_{Fi} in Eq. 5 by $ID_{Fi,b}$ (information defect of Fi after balancing) and it is $U_b \leq U_{ap}$. As both U_b and U_{ap} are composites of S, the difference between the actual system uncertainty (U_{ap} or U_b) and the system size S is referred to as the information content of a material flow system.

<heading level 2> Weighted uncertainty of balanced material flow systems

While some flow quantities X_{Fi} are known before balancing (*a priori* data with $ID_{Fi} \epsilon (0,1)$), others are typically unknown (data gaps with ID_{Fi}=1). After balancing a system, all flow quantities $X_{Fi,b}$ in a system are known. Some of these flows may be quantitatively more relevant than others. Intuitively, knowing quantitatively major flows better contributes more to the overall state of knowledge about a material flow system than knowing quantitatively minor flows better. To also consider the quantitative relevance of flows within a system, the uncertainty measure U is adapted. Each flow is weighted by $\frac{X_{Fi,b}}{\sum_i X_{Fi,b}}$, where $X_{Fi,b}$ is the quantity of a balanced flow, multiplied with the number of flows n_F so as not to change the magnitude

of the summed *U*-measure. Combining the balanced information defects and the balanced flow quantities, this gives the weighted uncertainty measure $U_{b,w}$, which is

$$U_{b,w} = -\sum_{i=1}^{n_F} \frac{X_{Fi,b} n_F}{\sum_i X_{Fi,b}} \, ID_{Fi,b} \, log \frac{ID_{Fi,b}}{\sum_i X_{Fi,b}} \qquad \text{Eq. 6}$$

By means of Eq. 6, it can be expressed that, given data of good quality on the quantitatively most relevant flows of a system (large $X_{Fi,b}$), the information content of the system is high or, conversely, the uncertainty of this system is low.

<heading level 1> Complexity of material flow systems

As motivated by Allenby (2009), complexity may be regarded as a system property which involves both the system size and linkage patterns within a system. This relates to the understanding of Rutledge and colleagues, where a system is maximally complex (or non-trivial) if it consists of many elements and when each of these elements is connected to every other element in the system. In contrast, a trivial network structure such as a line network is of little or no complexity, even though it may be of considerable size. Recalling the useful feature of Eq. 2 and Eq. 3 to allow for combined consideration of intensive (system structure) and extensive (system size) dimensions motivates to apply the metrics for aggregated characterization of MFA networks.

Each flow *Fi* is considered to define a subset. In an MFA system, each flow *Fi* connects a source process y_i to a target process z_i. At both its source and target process, *Fi* probably has a number of neighboring flows, that is, flows that either also originate from y_i or that also enter z_i. Referring to the number of flows leaving y_i and entering z_i as the outdegree of process y_i (n_{yi}) and the indegree of process z_i (n_{zi}), denoting a flow *Fi* from y_i to z_i as n_{yizi}=1 and considering the total number of flows n_F in a system, Eq. 2 is reformulated as a measure for the triviality *T* of a system, so that

$$T = n_F \sum_{i=1}^{n_F} \frac{n_{yizi}}{n_F} log \frac{n_{yizi} n_F}{n_{yi} n_{zi}} = \sum_{i=1}^{n_F} \log \frac{n_F}{n_{yi} n_{zi}} \qquad \text{Eq. 7}$$

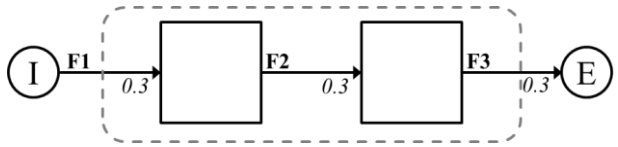
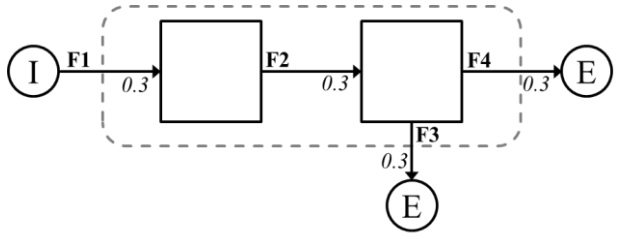
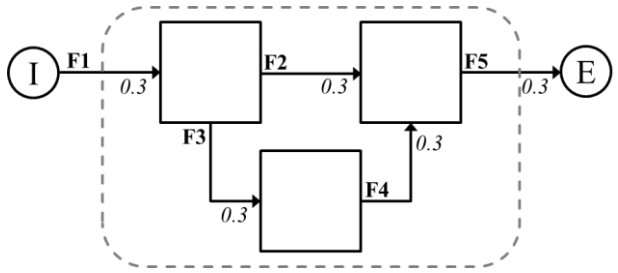
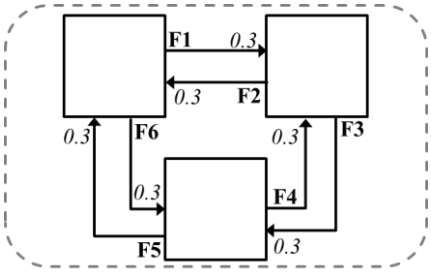
and its counterpart, the complexity *C* of a system, is formulated as

$$C = -n_F \sum_{i=1}^{n_F} \frac{n_{yizi}}{n_F} log \frac{n_{yizi}^2}{n_{yi} n_{zi}} = -\sum_{i=1}^{n_F} \log \frac{1}{n_{yi} n_{zi}} \qquad \text{Eq. 8}$$

Simple examples are provided in Table 1. In the most trivial topology (a line network, example A), it is *S*=*T* and *C*=0. *C* increases with n_F and more complicated linkage patterns in

example B and example C. Under the condition that no process connects to any other process by more than one flow, systems with n_p processes are maximally complex if they have $n_{F,max}=(n_p\text{-}1)\cdot n_p$ flows and if each process connects to every other process in the system (example D). In such maximally complex topologies, there always is a *T* component and, with increasing $n_{F,max}$, it is $\frac{C}{S} \to 1$ and $\frac{T}{S} \to 0$. For all above described topologies, it is $S=T+C$.

Table 1: Four examples (A-D) for illustration of the proposed system measures S (system size), U (uncertainty), T (triviality) and C (complexity). F1-F5 are the flow numbers. The numbers next to the flows designate the information defects ID_{Fi} (here considered equal for all flows to illustrate the influence of system size on the U measures). The dotted line represents the system boundary, flows crossing the system boundary are referred to as import or export flows.

	Graph	System measures
A		S = 4.8 U = 1.4 T = 4.8 C = 0.0
B		S = 8.0 U = 2.4 T = 6.0 C = 2.0
C		S = 11.6 U = 3.5 T = 7.6 C = 4.0
D		S = 15.6 U = 4.7 T = 3.6 C = 12.0

Real-world MFAs typically are between the extreme cases illustrated in example A and example D in Table 1. As line networks are untypical topologies of material flow systems, there always is a *C* component in a realistic MFA. Because of MFA-specific structural limitations, *C* is never maximal. These limitations include that flows crossing the system

boundary originate from or enter processes with an outdegree (in the case of import flows) or an indegree (in the case of export flows) of one. Also, pairs of processes are typically not connected in both directions but in one direction only. According to Eq. 4 - Eq. 8, each individual flow *Fi* contributes to the aggregated measures by a specific degree, which enables distinguishing flows from one another according to their respective uncertainty in the system context and their configuration in the system structure. This is further elaborated in two case studies presented in the following.

<heading level 1> Analysis and comparison of two material flow systems

The application of the proposed measures has been illustrated in simple hypothetical examples in Table 1. They can also be applied to full-scale MFAs, as presented in this section.

Buchner et al. (2014) provide a detailed analysis of aluminum (Al) flows in Austria for the year 2010 (Figure 1). The aluminum MFA consists of 77 flows (n_F=77). The sum of all $X_{Fi,b}$ in the system is about 4600 kilotonnes per year.

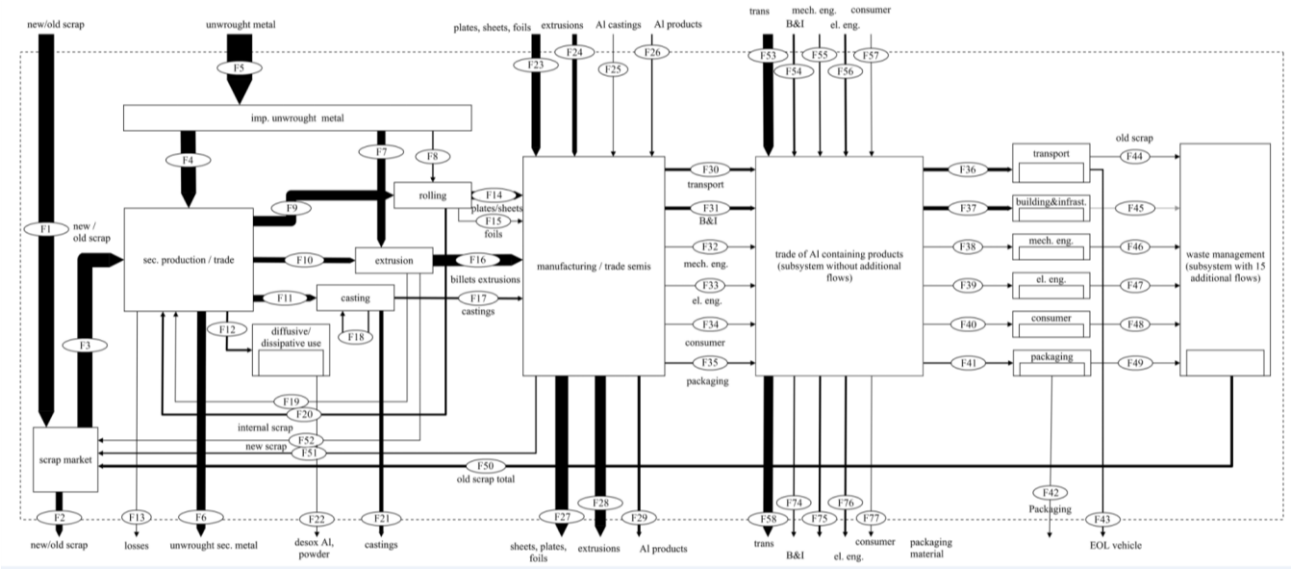


Figure 1: Flowchart of the 2010 Austrian aluminum flow system (Buchner et al. 2014).

In van Eygen et al. (2016) a detailed study of plastics flows in Austria for the year 2010 is presented (Figure 2). The plastics MFA consists of 88 flows (n_F=88). The sum of all $X_{Fi,b}$ in the system is about 15,000 kilotonnes per year.

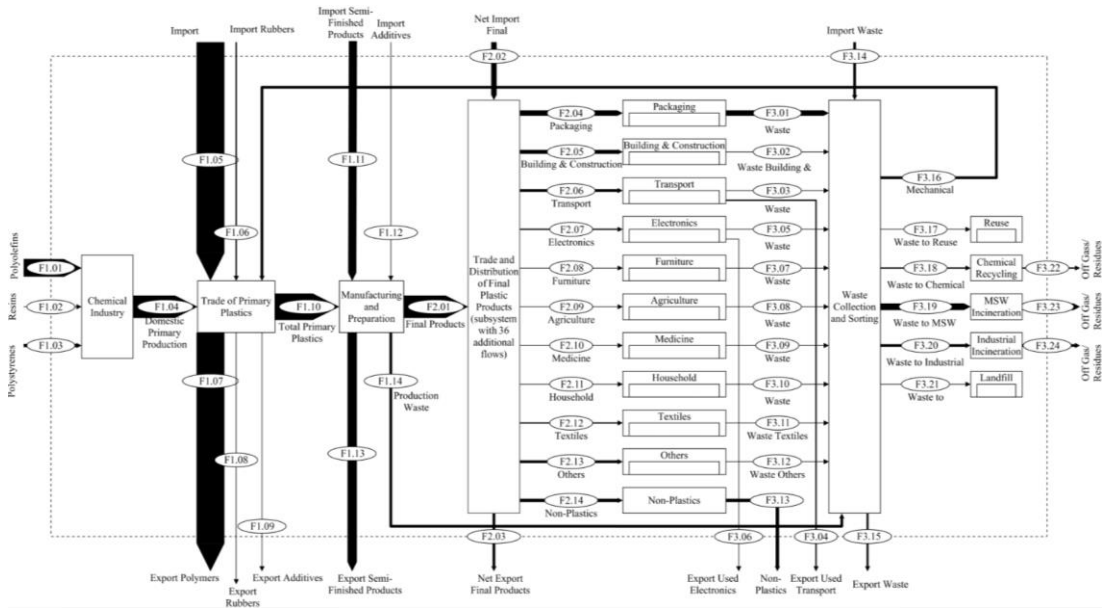


Figure 2: Flowchart of the 2010 Austrian plastics flow system (van Eygen at al 2016).

The quality of model input data (information defects ID_{Fi}) of the two studies (Figure 1, Figure 2) has been evaluated according to Schwab et al. (2016a). The measures for uncertainty and complexity in MFA (calculated according to Eq. 4- Eq. 8) of the case studies are listed in Table 2. A list of all input variables and of each individual flows' contribution to the total system uncertainty and complexity is provided in appendix S1.

Table 2: System measures of the aluminum and plastics systems (rounded to the nearest integer)

	Aluminum	Plastics
System size		
S	483	568
Uncertainty		
U_{ap}	196	335
U_b	99	202
$U_{b.w}$	71	168
Structure		
T	270	296
C	212	272

The absolute magnitude of the measures enables comparing the aluminum and the plastics systems to one another in an absolute sense. The fact that the Al system consists of fewer flows than the plastics system reflects in the measure S (Table 2). For systems of larger or

smaller size, *S* differs more significantly: Applied to the Austrian 2009 Phosphorus MFA (Zoboli et al. 2015; Schwab et al. 2016b), a comparatively detailed system with 122 flows, it is *S*= 846. For a palladium MFA with 25 flows (Laner et al. 2015a; Schwab et al. 2016a) it is *S*= 116. In addition to their absolute magnitude, relative comparisons of the measures provide useful information about MFA systems, as illustrated in Figure 3.

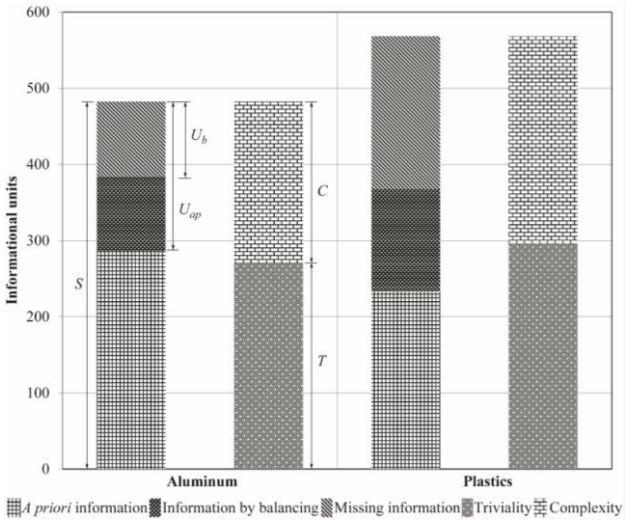


Figure 3: Information content of the aluminum and plastics MFAs (calculated as differences between S, U_{ap} and U_b) and their structural triviality and complexity.

By means of the ratio of the measures *S*, U_{ap} and U_b, the information gain from a qualitative system to a system with *a priori* information and a balanced system is quantified. The information content increases with decreasing uncertainty in the system (Figure 3). The initial uncertainty of the qualitative Al system is *S*=483. Considering the available *a priori* information on flows *Fi*, the uncertainty decreases to U_{ap}= 196. By balancing (data reconciliation and bridging of data gaps), the uncertainty in the system decreases to U_b= 99. In the plastics system, the uncertainty decreases from *S*=568 to U_{ap}=335 and to U_b=202. The plastics flow system is both relatively and absolutely speaking more complex than the aluminum flow system, as it can be seen when comparing the absolute magnitude of *C* and the relation of *C* to *S* of the both case studies. In both systems, $U_{b,w}$ is lower than U_b. This indicates that, both in the Al and plastics system, quantitatively dominating flows are known better than quantitatively minor flows (Table 2). The *a priori* data of the Al system is considerably better than the *a priori* data of the plastics system. This is indicated by the fact that U_{ap} equals less than half of *S* in the Al system, while U_{ap} of the plastics system is only two fifth lower than *S*. By balancing, the uncertainty of the aluminum system decreases by 21% and the uncertainty of the plastics system by 24%. The relation of $U_{b,w}$ to *S* indicates that

the aluminum system is known to an extent of 85% and the plastics system is known to an extent of 70%.

<heading level 1> Usefulness and limitations

A convenient feature of the uncertainty and complexity measures presented in this article is that both phenomena are quantified as formally linked measures and expressed in the same abstract dimension. They enable evaluating the system structure and the state of knowledge on different MFAs or on MFAs at different points in their development process. The information gained by performing MFA procedures can be quantified and compared. As shown in the aluminum and plastics case studies, the information content of material flow systems can instantly be derived from the uncertainty of systems once this uncertainty is quantified.

Transparent and informative communication of results has been observed to be a shortcoming in different fields of Industrial Ecology (Lazarevic et al. 2012). This holds also for MFA. The proposed measures are one possibility to help transporting information not only on actual material quantities, but also on the reliability of model results. For making well-informed decisions for example in resource management, it may be relevant, in addition to knowing about aggregate system properties, to know about particularly certain or uncertain flows or sectors in the system. This can be represented in a convenient way by means of flowcharts, as proposed in Figure 4.

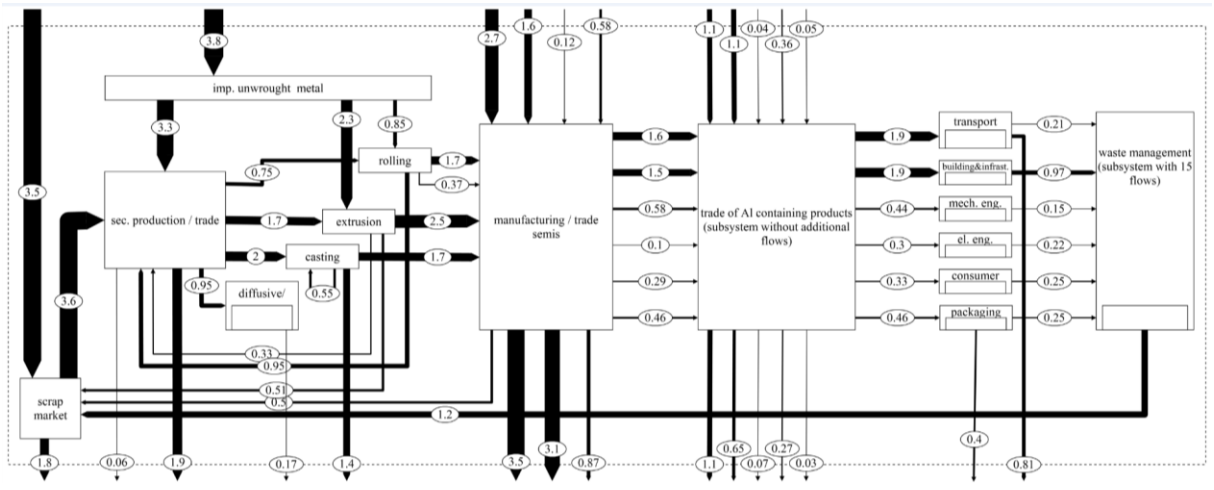


Figure 4: Uncertainty $U_{b,w}$ of the flows in the aluminum MFA displayed as a flowchart. Flow widths are proportional to the weighted uncertainty per flow (Eq. 6). For the total system, it is $U_{b,w}=71$.

Comparing individual flows of the same system by their contribution to $U_{b,w}$ (that is, by the components of the sum in Eq. 6) reveals relevant information for identifying critical flows,

weaknesses in the system and needs for further investigations, which is relevant information for both MFA modelers and decision makers. In the aluminum system, quantitatively major flows contribute most to system uncertainty $U_{b,w}$, even though they are mostly known better than quantitatively minor flows (compare Figure 1 and Figure 4 and see also appendix S1 in the supporting information on the web). Further visualizations of uncertainty *U* by means of flowcharts are provided in appendix S2 in the supporting information on the web.

A practical characteristic of the measures *S* and *U* as presented in this article is that it represents both the information needed for construction of a qualitative system of flows and the data for quantification of these flows. The more flows there are in a system, the higher is the total system uncertainty (resp., the potential information content of a system) and the more and better data are needed to minimize *U*. In an informationally optimized system, it is $(S-U_{ap})/S \rightarrow 1$. An optimum is reached by increasing *S* (distinguishing more flows) while decreasing U_{ap} (incorporating better data so that $U_{ap} \rightarrow 0$). This antagonistic interpretation of *S* and U_{ap} shows that, depending on the available information basis, it is usually not helpful to increase the resolution of a system when no adequate data is provided. In return, given rich and detailed information, increasing the system resolution is beneficial in the sense of increasing *S* and decreasing U_{ap}.

The difference between U_{ap} and U_b indicates the amount of information gained by system balancing, and the degree to which *a priori* data of an MFA system is conflicting and has to be manipulated to meet mass balance constraints (data reconciliation). A high degree of data reconciliation indicates either that the available material flow data is inconsistent, or that the qualitative system is unrealistic or incomplete and has to be revised (or both).

It has to be recalled that the actual design of a system may for various reasons not be complete and representative, for example because it depends on probably arbitrary choices of agents (Heijungs 2015). Although differences in system design have been identified as to be relevant in MFA, distinctions and comparisons regarding system structure are intricate (Klinglmair et al. 2016). The proposed measures *T* and *C* provide an improved basis for future analyses and comparisons. While the term "complexity" of MFAs has been used in a qualitative manner (Klinglmair et al. 2016), it can now be expressed by quantitative means. Moreover, flows that contribute most to the complexity of a system and sectors of particularly high or low complexity can be identified and compared quantitatively (Figure 5).

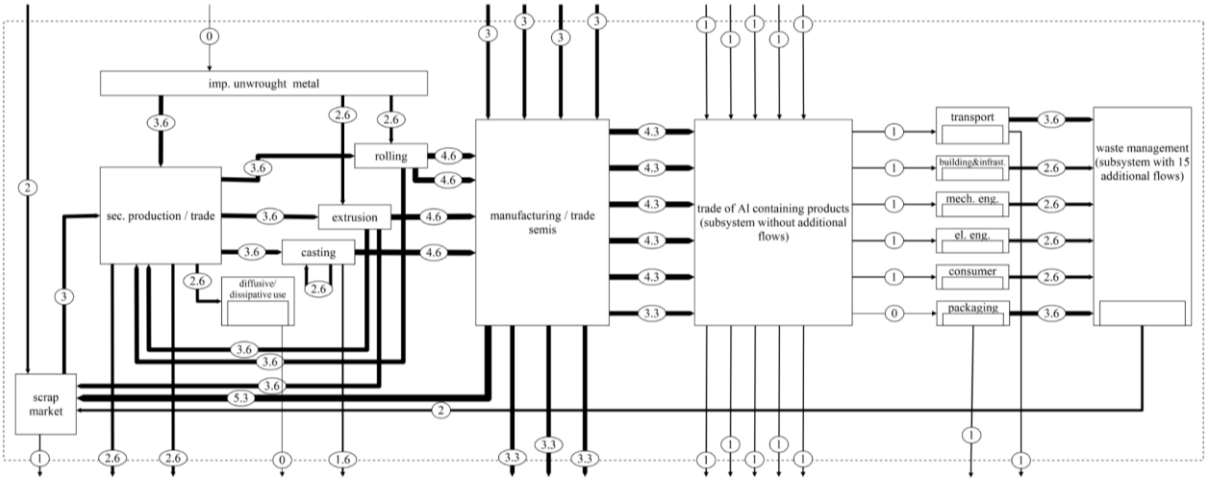


Figure 5: Complexity C of the flows in the aluminum MFA displayed as a flowchart. Flow widths are proportional to the contribution per flow to C (Eq. 8). For the total system, it is C=212.

A flowchart on the complexity of the plastics system is provided in appendix S2 in the supporting information on the web. Increasing complexity *C* of a system relates to increasing information demand of this system. Practically speaking, this means that an agents' effort for data acquisition increases with *C*. Visualizing *C* as a flowchart (Figure 5), flows that contribute most to the complexity *C* can be identified, and these are the flows that are most important to be known *a priori* in order to balance a system.

The proposed procedure for structural analysis is limited regarding adequate representation of some MFA-untypical topologies, which may be chosen by agents performing an analysis. This applies for example to loops (flows leaving and entering the same process in the studied unit time interval), such as flow 18 in the aluminum case study, and to parallel flows (flows from and to the same source and target process), such as flows 14 and 15 in the aluminum case study (Figure 1). Parallel flows result from flows being represented in a more disaggregated manner than processes and can in virtually all cases be prevented by respective system design, that is, by disaggregating either the source process or the target process, for example by use of subsystems (as it is for flows 30-35 in Figure 1). For one topological particularity, the characteristic of the structural measures that the contribution of each *Fi* to *C* and *T* is expressed in relation to the total number of flows in the system yields notable, though unproblematic, results. For a *Fi* that connects processes with particularly high outdegree n_{yi} and indegree n_{zi}, so that $n_{yi} \cdot n_{zi} > n_F$, the contribution of this *Fi* to *T* takes negative values (cf. the concept of "pointwise mutual information" in information theory). This is not the case for any flow in the two systems analyzed in this article and, though conceptually possible, improbable in real-world MFAs. Such situations reflect particularly high structural heterogeneity of systems, which entails that the proportions different flows have in *C* and *T*

shift. A *Fi* with a negative contribution to *T* has a proportionately higher contribution to *C* and the aggregate measures *C* and *T* sum up to *S*.

The procedures presented in this study focuses on flows. Transfer coefficients, stocks and stock change rates, which are also entities that introduce information or uncertainty into material flow systems and that add to the complexity of systems, are not considered. It thus may be useful to implement these entities into the measures in further research. Stock change rates, for example, can be treated identically to import or export flows (which originate from or enter processes with outdegree or indegree one). That way, they would contribute to *S*, reflect in the structural measures *T* and *C* and, after assigning an *ID*, also in the uncertainty measures *U*.

The proposed procedures may also be applied to other tools in the field of Industrial Ecology. In principle, they can be used for analysis of all systems that can be represented as networks, such as life cycle inventories or input-output models. In MFA, they complement existing methodologies by supporting system design, optimized use of available information and communication of MFA results. Issues of data quality and system structure, which have been qualitatively discussed for example by Klinglmair et al. (2016) can now, by means of the measures proposed in this article, be gauged and quantitatively compared. Despite these possibilities, information content and system design in MFA are, in the presence of limited information, inescapably subjective to a certain degree. This is both a limitation and an incentive of the procedure proposed in this thesis, which makes it formally possible to work with limited information in a transparent way.

<heading level 1> Nomenclature

Abbreviations

C	Complexity
Fi	Flow i
ID_{Fi}	Information defect of Fi (subscript b – balanced)
k	Positive constant
n_F	Number of flows in system
S	System size
T	Triviality
U	System uncertainty (subscripts ap, b, w – a priori, balanced, weighted)
X_{Fi}	Quantity of a flow (subscript b – balanced)

y Source process

z Target process

<heading level 1> Acknowledgements

The authors thank Nađa Džubur and Oliver Cencic (both TU Wien) for helpful comments to this research, and two anonymous reviewers for their useful comments to a previous version of this article. This research was funded by the Austrian Federal Ministry of Science, Research and Economy.

Supporting Information

Appendix S1: Input and output variables per flow for calculation of uncertainty and complexity of the aluminum and plastics case studies

Appendix S2: Additional flowcharts visualizing the uncertainty and complexity of the aluminum and plastics systems

<heading level 1> References

Allenby, B. 2009. The industrial ecology of emerging technologies. *Journal of Industrial Ecology* 13(2): 168-183.

Bailey, R., B. Bras, and J. K. Allen. 2004. Applying Ecological Input-Output Flow Analysis to Material Flows in Industrial Systems: Part II: Flow Metrics. *Journal of Industrial Ecology* 8(1-2): 69-91.

Baird, D. and R. E. Ulanowicz. 1989. The Seasonal Dynamics of The Chesapeake Bay Ecosystem. *Ecological Monographs* 59(4): 329-364.

Bettencourt, L. M. A. and C. Brelsford. 2015. Industrial Ecology: The View From Complex Systems. *Journal of Industrial Ecology* 19(2): 195-197.

Brunner, P. H. and H. Rechberger. 2016. *Handbook of Material Flow Analysis: For Environmental, Resource, and Waste Engineers, Second Edition*: CRC Press.

Buchner, H., D. Laner, H. Rechberger, and J. Fellner. 2014. In-depth analysis of aluminum flows in Austria as a basis to increase resource efficiency. *Resources, Conservation and Recycling* 93(0): 112-123.

Cencic, O. 2016. Nonlinear Data reconciliation in Material Flow Analysis with Software STAN. *Sustainable Environment Research*.

Cencic, O. and H. Rechberger. 2008. Material Flow Anaylsis with Software STAN. *Journal of Environmental Engineering and Management* 18(1): 3-7.

Chen, W.-Q. and T. E. Graedel. 2012. Anthropogenic Cycles of the Elements: A Critical Review. *Environmental Science & Technology* 46(16): 8574-8586.

Christian, R. R., D. Baird, J. Luczkovich, J. C. Johnson, U. M. Scharler, and R. E. Ulanowicz. 2005. Role of network analysis in comparative ecosystem ecology of estuaries. *Aquatic Food Webs* 3: 25e40.

Côté, R. and J. Hall. 1995. Industrial parks as ecosystems. *Journal of Cleaner Production* 3(1–2): 41-46.

Dijkema, G. P. J. and L. Basson. 2009. Complexity and Industrial Ecology. *Journal of Industrial Ecology* 13(2): 157-164.

Dijkema, G. P. J., M. Xu, S. Derrible, and R. Lifset. 2015. Complexity in Industrial Ecology: Models, Analysis, and Actions. *Journal of Industrial Ecology* 19(2): 189-194.

Goerner, S. J., B. Lietaer, and R. E. Ulanowicz. 2009. Quantifying economic sustainability: Implications for free-enterprise theory, policy and practice. *Ecological Economics* 69(1): 76-81.

Gottschalk, F., R. W. Scholz, and B. Nowack. 2010. Probabilistic material flow modeling for assessing the environmental exposure to compounds: Methodology and an application to engineered nano-TiO2 particles. *Environmental Modelling & Software* 25(3): 320-332.

Graedel, T. E. 1996. On the concept of Industrial Ecology. *Annual Review of Energy and the Environment* 21(1): 69-98.

Graedel, T. E., D. van Beers, M. Bertram, K. Fuse, R. B. Gordon, A. Gritsinin, A. Kapur, R. J. Klee, R. J. Lifset, L. Memon, H. Rechberger, S. Spatari, and D. Vexler. 2004. Multilevel Cycle of Anthropogenic Copper. *Environmental Science & Technology* 38(4): 1242-1252.

Hedbrant, J. and L. Sörme. 2001. Data Vagueness and Uncertainties in Urban Heavy-Metal Data Collection. *Water, Air and Soil Pollution: Focus* 1(3-4): 43-53.

Heijungs, R. 2015. Topological network theory and its application to LCA and IOA and related industrial ecology tools: principles and promise. *J Environ Account Manag* 3(2): 151-167.

Huang, D.-B., H.-P. Bader, R. Scheidegger, R. Schertenleib, and W. Gujer. 2007. Confronting limitations: New solutions required for urban water management in Kunming City. *Journal of Environmental Management* 84(1): 49-61.

Kay, J. J. 2002. On complexity theory, exergy and industrial ecology. In *Construction ecology - Nature as the basis for green buildings*, edited by C. J. Kibert, et al. New York: Spon Press.

Kharrazi, A., E. Rovenskaya, B. D. Fath, M. Yarime, and S. Kraines. 2013. Quantifying the sustainability of economic resource networks: An ecological information-based approach. *Ecological Economics* 90: 177-186.

Klinglmair, M., O. Zoboli, D. Laner, H. Rechberger, T. F. Astrup, and C. Scheutz. 2016. The effect of data structure and model choices on MFA results: A comparison of phosphorus balances for Denmark and Austria. *Resources, Conservation and Recycling* 109: 166-175.

Kopec, G. M., J. M. Allwood, J. M. Cullen, and D. Ralph. 2015. A General Nonlinear Least Squares Data Reconciliation and Estimation Method for Material Flow Analysis. *Journal of Industrial Ecology*: n/a-n/a.

Korhonen, J. 2001. Four ecosystem principles for an industrial ecosystem. *Journal of Cleaner Production* 9(3): 253-259.

Laner, D., H. Rechberger, and T. Astrup. 2014. Systematic Evaluation of Uncertainty in Material Flow Analysis. *Journal of Industrial Ecology* 18(6).

Laner, D., H. Rechberger, and T. Astrup. 2015a. Applying fuzzy and probabilistic uncertainty concepts to the material flow analysis of palladium in Austria. *Journal of Industrial Ecology* 19(6): 1055-1069.

Laner, D., J. Feketitsch, H. Rechberger, and J. Fellner. 2015b. A novel approach to characterize data uncertainty in MFA and its applications to plastic flows in Austria. *Journal of Industrial Ecology*.

Layton, A., B. Bras, and M. Weissburg. 2016. Industrial Ecosystems and Food Webs: An Expansion and Update of Existing Data for Eco-Industrial Parks and Understanding the Ecological Food Webs They Wish to Mimic. *Journal of Industrial Ecology* 20(1): 85-98.

Lazarevic, D., N. Buclet, and N. Brandt. 2012. The application of life cycle thinking in the context of European waste policy. *Journal of Cleaner Production* 29–30: 199-207.

Lenzen, M., R. Wood, and T. Wiedmann. 2010. Uncertainty Analysis for Multi-Region Input-Output Models - A Case Study of the UK's Carbon Footprint. *Economic Systems Research* 22(1): 43-63.

Lloyd, S. M. and R. Ries. 2007. Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment: A Survey of Quantitative Approaches. *Journal of Industrial Ecology* 11(1): 161-179.

Meerow, S. and J. P. Newell. 2015. Resilience and Complexity: A Bibliometric Review and Prospects for Industrial Ecology. *Journal of Industrial Ecology* 19(2): 236-251.

Morf, L. S. and P. H. Brunner. 1998. The MSW Incinerator as a Monitoring Tool for Waste Management. *Environmental Science & Technology* 32(12): 1825-1831.

Navarrete-Gutiérrez, T., B. Rugani, Y. Pigné, A. Marvuglia, and E. Benetto. 2015. On the Complexity of Life Cycle Inventory Networks: Role of Life Cycle Processes with Network Analysis. *Journal of Industrial Ecology*: n/a-n/a.

Nuss, P., W.-Q. Chen, H. Ohno, and T. E. Graedel. 2016. Structural Investigation of Aluminum in the U.S. Economy using Network Analysis. *Environmental Science & Technology* 50(7): 4091-4101.

Odum, H. T. 1994. *Ecological and General Systems: An Introduction to Systems Ecology*: University Press of Colorado.

Ott, C. and H. Rechberger. 2012. The European phosphorus balance. *Resources, Conservation and Recycling* 60: 159-172.

Rutledge, R. W., B. L. Basore, and R. J. Mulholland. 1976. Ecological stability: An information theory viewpoint. *Journal of Theoretical Biology* 57(2): 355-371.

Schiller, F., A. S. Penn, and L. Basson. 2014. Analyzing networks in industrial ecology – a review of Social-Material Network Analyses. *Journal of Cleaner Production* 76: 1-11.

Schwab, O., D. Laner, and H. Rechberger. 2016a. Quantitative evaluation of data quality in regional Material Flow Analysis. *Journal of Industrial Ecology*.

Schwab, O., O. Zoboli, and H. Rechberger. 2016b. A Data Characterization Framework for Material Flow Analysis. *Journal of Industrial Ecology*.

Suh, S. 2005. Theory of materials and energy flow analysis in ecology and economics. *Ecological Modelling* 189(3–4): 251-269.

Trinkel, V., N. Kieberger, T. Bürgler, H. Rechberger, and J. Fellner. 2015. Influence of waste plastic utilisation in blast furnace on heavy metal emissions. *Journal of Cleaner Production* 94: 312-320.

Ulanowicz, R. E. 1980. An hypothesis on the development of natural communities. *Journal of Theoretical Biology* 85(2): 223-245.

Ulanowicz, R. E. 1997. *Ecology, the Ascendent Perspective*: Columbia University Press.

Ulanowicz, R. E., S. J. Goerner, B. Lietaer, and R. Gomez. 2009. Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity* 6(1): 27-36.

van Eygen, E., J. Feketitsch, D. Laner, H. Rechberger, and J. Fellner. 2016. Comprehensive analysis and quantification of national plastic flows: the case of Austria. *Resources, Conservation and Recycling*.

Velis, C. A., S. Wagland, P. Longhurst, B. Robson, K. Sinfield, S. Wise, and S. Pollard. 2013. Solid Recovered Fuel: Materials Flow Analysis and Fuel Property Development during the Mechanical Processing of Biodried Waste. *Environmental Science & Technology* 47(6): 2957-2965.

Weidema, B. P. and M. S. Wesnæs. 1996. Data quality management for life cycle inventories—an example of using data quality indicators. *Journal of Cleaner Production* 4(3–4): 167-174.

Wood, R. and M. Lenzen. 2009. Aggregate Measures of Complex Economic Structure and Evolution. *Journal of Industrial Ecology* 13(2): 264-283.

Wulff, F., J. G. Field, and K. H. Mann. 1989. *Network analysis in Marine Ecology: Methods and Applications.*: Springer.

Zoboli, O., M. Zessner, and H. Rechberger. 2015. Added Value of Time Series in MFA: The Austrian Phosphorus Budget from 1990 to 2011. *Journal of Industrial Ecology*.

Zoboli, O., M. Zessner, and H. Rechberger. 2016. Supporting phosphorus management in Austria: Potential, priorities and limitations. *Science of the total Environment* 565: 313-323.