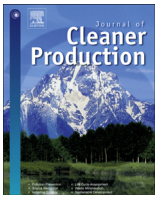
Journal of Cleaner Production 137 (2016) 1109–1117

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro



Semantic catalogs for life cycle assessment data

Brandon Kuczenski [a, *], Christopher B. Davis [b], Beatriz Rivela [c, d], Krzysztof Janowicz [e]

[a] *Institute for Social, Behavioral, and Economic Research, University of California, Santa Barbara, Santa Barbara, CA 93106-2150, USA*
[b] *Faculty of Mathematics and Natural Sciences, University of Groningen, Groningen, The Netherlands*
[c] *inViable Life Cycle Thinking, Madrid, Spain*
[d] *Institute of Habitat Science, National Polytechnic School, Quito, Ecuador*
[e] *Department of Geography, University of California, Santa Barbara, Santa Barbara, CA USA*

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form
23 July 2016
Accepted 31 July 2016
Available online 2 August 2016

Keywords:
Life cycle assessment
Interoperability
Semantic web
Ontology design patterns
Data management
Linked data

ABSTRACT

Life cycle assessment (LCA) is a highly interdisciplinary field that requires knowledge from different domains to be gathered and interpreted together. Although there are relatively few major data sources for LCA, the data themselves are presented with highly heterogeneous formats, interfaces, and distribution mechanisms. The lack of agreement among data providers for descriptions of processes and flows creates substantial barriers for information sharing and reuse of practitioners' models.

Nevertheless, the many data resources share a common logic. The use of Semantic Web technologies and text mining techniques can facilitate the interpretation of data from diverse sources. Numerous existing efforts have been made to articulate a knowledge model for LCA. In March of 2015 a joint workshop was held that brought together leading international domain experts with ontology engineers to develop a set of simple models called ontology design patterns (ODPs) for LCA information. In this paper we build on the outcomes of the workshop, as well as prior published works, to derive a minimal "consensus model" for LCA. We use the consensus model to derive a description of an LCA "catalog" that can be used to express the semantic content of a data resource. We generate catalogs of several prominent databases, and make those catalogs available to the public for independent use. Finally, we "link" those catalogs to existing knowledge models using JSON-LD, a linked data format that can expose the catalog contents to Semantic Web tools.

We then show by example how the catalogs may be used to answer questions about the scope, coverage, and comparability of data, both within and across data sources, that are difficult to answer when the contents of the catalogs are provided independently and inconsistently. We discuss how the use of semantic catalogs can help address challenges that initiatives such as the "Global Network of Interoperable LCA Databases – Global LCA Data Access" are facing today.

1. Introduction

1.1. The challenges of an interdisciplinary field

Achieving sustainable production and consumption requires coordinated efforts across disciplinary boundaries from public agencies, industry actors and researchers around the world. The Millennium Development Goals of the 1992 UN Conference on Environment and Development have been formally superseded by the 17 Sustainable Development Goals (SDGs) and their related 169 targets (UN, 2015; UN, 2014), launched in September 2015. All the components of sustainability are required to be implemented, posing an unprecedented challenge to policy makers and scientists. It is evident that scientific, quantitative sustainability assessments are the key to monitoring the progress of and support the decisions on the 17 SDGs (Hák et al., 2016). Despite scientific progress in individual disciplines, it remains difficult to synthesize the efforts of multiple groups to meet global challenges, such as combating climate change while ensuring energy access for all (Cucurachi and Suh, 2015).

Quantitative sustainability assessments, and life cycle assessment (LCA) in particular, cannot be completed without integrating information from diverse areas of knowledge. Many users from the academia, government, industry, and consultancies all over the

* Corresponding author.
E-mail address: bkuczenski@ucsb.edu (B. Kuczenski).

http://dx.doi.org/10.1016/j.jclepro.2016.07.216

world apply LCA in scientific, industrial, agricultural, societal, or political processes, use their findings to evaluate and improve current practices (Baitz et al., 2013), and are proactively working in methodology development, data provision, data curation, or product optimization and communication. The variety of data formats, storage formats, system definitions, and software implementations has been demonstrated to be a major problem for LCA (Speck et al., 2015; Ingwersen, 2015; Ingwersen et al., 2015a; Herrmann and Moltesen, 2015). The SETAC-Europe LCA Working Group “Data Availability and Data Quality” focused on a common data exchange format, public databases and accepted quality measures (Hischier et al., 2001), concluding that more rigid formulations were impractical or counterproductive. The publication ‘Global Guidance Principles for Life Cycle Assessment (LCA) Databases’ (Sonnemann et al., 2011) illustrates both the progress made in detailing common principles and the difficulties presented by diverging visions. International efforts to promote consistency and interoperability have culminated in the launch of the Global Network of interoperable LCA Databases (in 2014), an initiative from the International Forum on LCA Cooperation with the vision to establish a global network comprised of independently-operated data resources (Mila i Canals et al., 2015).

Although the inadequacy of current techniques for managing LCA data and computations has been discussed in one form or another for many years (Ayres, 1995; Owens, 1997), a gulf has remained between high-level guidance initiatives and common practice. Existing approaches, such as the widely-used Ecospold and ILCD exchange formats, mainly address syntactic interoperability and typically fall short of approaching the underlying problems of dealing with heterogenous data, namely differences in semantics.

As an example, consider Table 1, which shows information pertaining to the fuel economy of truck transport in the US LCI and Ecoinvent databases. Although the systems being modeled are similar, textual information describing the data set properties is widely varying. The metadata include units of measure, geographic boundaries, and time frames, as well as synonymous terms (e.g. “truck” versus “lorry”), all of which require different types of knowledge to interpret. Though the data sets are technically interoperable, what is ultimately needed are tools that can support users in interpreting the data and evaluating the data sets' appropriateness for their specific applications.

In this paper, we present a “data-first” approach to the interoperability problem in the LCA field. We propose a “consensus model” of the core concepts in LCA, and use it to develop catalogs of several of the most prominent inventory data resources. We publish the catalogs using linked data technology that enables their contents to be automatically understood by Semantic Web tools. We show how the catalogs can be used to generate a wide range of insights about the contents, similarities and differences among data sources. The catalog approach demonstrates a path forward for improving the accessibility and interpretation of interoperable data resources.

2. Approach and methods

2.1. The promise of semantic methods

Managing vast quantities of information is made much easier with the support of automated tools for locating and interpreting data. The Semantic Web refers to data that can be automatically interpreted by machines (Bizer et al., 2009). There are two general requirements for producing machine-readable data: the *structure* of the data must be defined; and the *meaning* of the data must be formalized. The first requirement means allowing different data entities to “link” to one another, which is accomplished through the use of explicit web references, called URIs, IRIs, or hyperlinks. The second requirement is met by specifying the relationships between entities according to a formal representation of knowledge known as an ontology.

An ontology is a “formal model that uses mathematical logic to clarify and define concepts and relationships within a domain of interest” (Madin et al., 2008). In order for formal logic to be applied, an ontology is of necessity a precise, technically complex construction. A consensus-driven semantic model for socioeconomic metabolism (SEM) can support the development of practical data structures and databases (Pauliuk et al., 2015). More recently, Pauliuk et al. (2015a) have argued that the use of informal “practical ontologies” can facilitate semantic annotation of data and database development and thus help researchers to develop data infrastructure for SEM research. The use of ontology-based approaches for LCA was first suggested in 2005 (Kraines et al., 2005), where it was envisioned as part of a multi-tier system for distributed knowledge management in integrated environmental assessment. Such a system, when paired with Semantic Web services such as semantic search, was seen to be a great support for knowledge discovery (Kraines et al., 2006). Similar approaches are in development in the enterprise domain that could be adapted for distributed reasoning about sustainability (Muñoz et al., 2013). The use of Linked Open Data is critical to the success of such an enterprise (Davis et al., 2010), although the industrial ecology community has been slow to adopt this technology.

A number of groups have proposed ontologies for LCA. The open-source software Earthster was developed as an LCA-specific Linked Data application, culminating in the development of the Earthster Core Ontology (ECO) (Epimorphics Ltd, 2010; Sayan, 2011), but it failed to achieve significant community support and development halted in 2011. Independent efforts generated a reference semantic implementation of the US Life Cycle Inventory database (Bertin et al., 2012), a semantically-enriched model of refinery operations (Takhom et al., 2013), and an ontology for product manufacturing (Zhang et al., 2015). Other experiences and achievements regarding model and data harmonization have taken place within input–output frameworks (Lenzen et al., 2014). However, the various efforts do not share a common formal underpinning, which is an important condition for interoperability (Janowicz et al., 2014).

Table 1
Diesel truck transport in two databases.

Feature	Ecoinvent v3.2 Cut-off	US LCI
Process name	Transport, freight, lorry 16–32 metric ton, EURO5, cut-off, U (RoW)	Transport, combination truck, diesel powered
Spatiotemporal scope	RoW, 2009–2015	RNA, 2001-01-01
Functional unit	1 metric ton*km	1 t*km
Fuel flow name	Diesel, low-sulfur	Diesel, at refinery
Providing process	Market for diesel, low-sulfur, cut-off, U (RoW)	Petroleum refining, at refinery
Exchange value	0.03747 kg	0.0272 l

2.2. Community-driven ontology design – the vocabulary camp

Given the high interdisciplinarity and breadth of practice within LCA, arriving at a shared monolithic domain model does not seem like an attainable goal. Part of the power of the ontology-driven approach is that it does not require a universal agreement but instead merely requires that the formal knowledge model can be specified or inferred by the data context. The computational model of LCA forms the common ground that enables these mappings to be understood, although they were developed independently. An ontology design pattern (ODP) is a small, reusable knowledge model that is meant to be both precisely stated and also easy to adapt to different end-use situations (Gangemi, 2005).

One approach to developing an ODP is to generate use cases that capture recurring domain or cross-domain problems. These uses cases can guide the design of ontologies and help in its evaluation. The approach involves the use of competency questions (Grüninger and Fox, 1995); these are (often informal) queries that an ontology should be able to answer and that act as requirements for its axiomatization. While there is clear value to strong philosophical and deep domain approaches, the emphasis of this approach is clearly on utility.

This approach characterizes a series of modeling workshops known as "Vocabulary camps" or "GeoVoCamps"[1] that bring together ontology engineers with a selection of domain experts (Hitzler et al., 2015) in order to develop design patterns that address the needs of that domain. The products of the meeting are supposed to be immediately useable in a Semantic Web framework. The focus of the March 2015 meeting at the University of California, Santa Barbara (UCSB) was LCA and included the participation of an international group of LCA practitioners and scholars. The main outcomes from the workshop were further discussed and presented at the 14th International Semantic Web Conference (ISWC, 2015), in particular related to:

- An ontology pattern that specifies key aspects of LCA/LCI data models, i.e., the notions of flows, activities, agents, and products, as well as their properties (Janowicz et al., 2015).
- An ontology for modeling spatiotemporal scopes, i.e., the contexts in which inventory information or impact estimates are valid (Yan et al., 2015). Because environmental impacts always stem from industrial activities, the activities are the anchor points for spatial and temporal localization.

These design patterns were described in a formal logic, and reviewed and subsequently published by the ontology engineering community. A third ontology pattern pertaining to the characterization of environmental impacts was not completed and remains in progress.

2.3. The consensus model for LCA data

The ontologies and ODPs discussed above all have strengths and drawbacks that depend on the interests and experiences of the researchers who developed them. However, while none of them can be considered to be fully adequate, neither is any one of them incorrect. Nevertheless, because all the models are concerned with process-based LCA, they must have certain elements in common. In preparing this paper we reviewed existing knowledge models for LCA, including the above explicit semantic models, as well as the implicit models in established formats for LCI data exchange. Drawing on their commonalities, we develop a minimal "consensus model" that contains elements common to all.

The model is presented in Fig. 1. Three *classes* or *entity types* can be identified that are required by all models under development: "Activities," "Flows," and "Flow Quantities." The semantics of each of these entity types is complex, and each can ultimately be defined by an independent ontology. However, all of them are required by any conceptualization of LCA, and evidence of all three entity types can be found in every knowledge model studied. In simple terms, an Activity is a "thing that happens" and a Flow is a "thing in the world" that exists because of some instance of an Activity. A Flow has a direction with respect to an Activity: it is an output of one Activity and an input to another. The "Flow Quantity" represents a distinct quantitative characteristic that can be ascribed to a Flow.

Each instance of an entity class requires some external information in order to be well defined. Activities cannot be defined without knowledge of their spatiotemporal scope. Similarly, flows are not fully defined without knowing the "compartment" or medium that contains them. Finally, flow quantities must be defined in terms of some extensive unit of measure. Two relationships can be identified among class instances: "exchange" and "characterization." An exchange is an established relationship between an activity instance and a flow instance. Activities feature flows as inputs and outputs; meanwhile, each flow is an output of one activity and an input to another. Therefore, an exchange is sufficient to describe "half a flow instance," since one exchange specifies either the flow's origin or its terminus. In order to fully specify an exchange, it is necessary to specify a particular activity, a particular flow, and a direction, which is nominally "input" or "output".

Similarly, a "characterization" is an established relationship between a flow and a flow quantity. Flows typically have many characterizations. For instance, the flow of "gasoline" has mass, volume, economic value, toxicity potential, energy content, and others. As in the case of exchanges, the quantitative "value" of the characterization is not part of the semantic content of the relationship. A characterization has an implicit or explicit spatiotemporal scope which corresponds to the activity that generates or consumes the flow.

We emphasize that exchanges and characterizations in this model do not include quantitative information – they merely establish a relationship between entities. If a plastic forming requires an input of electricity to operate, that is sufficient to define the exchange relationship. The particular *exchange value,* that is, the quantity of electricity that is required in order to accomplish some task, is not part of the *semantic* relationship between the entities of "plastic forming process" and "electricity".

Computationally, LCIA indicators, such as estimates of global warming potential or toxicity potential, bear a strong similarity with other flow quantities, since they describe quantitative characteristics of specific flows. Some software systems already describe LCIA factors as "environmental quantities" that are similar to physical flow properties, although in the ILCD schema "Flow Properties" and "LCIA Methods" are distinct entity types. We suggest LCIA methods are equivalent to physical Flow Quantity entities, except having units of measure that describe potential environmental impacts.

2.4. Applying the consensus model to knowledge organization

From a knowledge modeling perspective, the notions of "flow", "activity", and "flow quantity" are *classes* – or abstract concepts. A *particular* flow, activity, or quantity is called an *entity* or an *instance* of a class. An LCA practitioner or data set developer creates a model by making observations of specific instances, not of the abstract concepts themselves. Exchanges and characterizations describe

[1] The Geo-prefix indicates the meetings' early focus on geographic information systems.

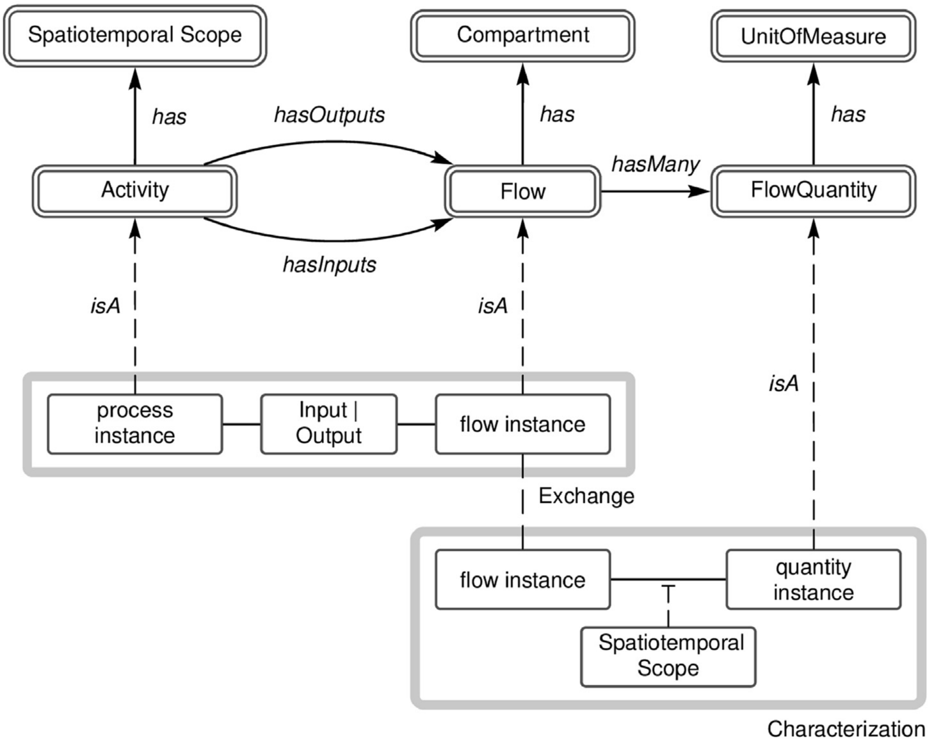


Fig. 1. The consensus knowledge model showing three entity types and their defining properties. An exchange reports a relationship between activity and flow instances, and a characterization reports a relationship between flow and flow quantity instances.

relationships among instances, not classes.

There is some variety in the LCA community regarding the use of the terms “activity” and “process” to refer to “things that happen.” ISO 14040, citing ISO 9000, defines a “process” as “set of interrelated or interacting activities that transforms inputs into outputs.” We interpret the ISO definition to refer to a specific instance of an activity that has observed input and output relationships, whereas an “activity” is a more general concept. In our interpretation, a process is an instance of an activity or a set of activities; only a process has an “inventory”; and only a process can have quantified exchange values.

With this in mind, it is possible to consider the consensus data model of Fig. 1 as a “lens” through which to view existing data resources. A particular inventory database is a collection of instances. The components of the data model make up the minimally necessary parts of a “catalog” describing the contents of the database. Namely, each collection has activities, flows, and quantities, and those entities are related to one another through exchanges and characterizations.

When used in this way, the model can facilitate the side-by-side comparison of resources with different formal constraints. Data sets represented in different serialization formats, such as the Ecospold format or the ILCD format, in spreadsheet models, and in web pages, can all be viewed from a common perspective. Because the model makes few “ontological commitments” and because those that it does make are universal to LCA (activities have spatial and temporal scopes; flows have compartments; quantities have units of measure), the model can be applied to existing data sets without compromising their embedded meaning. When multiple data resources are catalogued in the same way, it becomes straightforward to apply semantic techniques such as search, aggregation, text similarity algorithms, and others to all data resources.

Finally, in order to enable the model must be exposed to the machinery of the Semantic Web. Most of the concepts shown in Fig. 1 and their properties can be associated with classes that have already been defined in external ontologies. In the remainder of the paper we describe how the catalogs were created for several important LCI data sets and how they are linked to existing ontologies. We then use the catalogs to answer questions about the contents of the data sets and find similarities between them.

3. Results

To evaluate the utility of the consensus model as a framework for knowledge organization, we developed software to analyze four prominent life cycle inventory databases and express them as catalogs of the format discussed above. The following databases were studied:

- the US LCI database, provided in Ecospold v1 format from the LCA Commons;
- the GaBi professional database, 2016 edition, plus 22 extension databases, provided in ILCD format via the Thinkstep website;
- the European Life Cycle Initiative Database (ELCD), version 3.2, provided in ILCD format via the Joint Research Centre website;
- The ELCD implementation of 28 life cycle impact assessment methods with respect to ELCD elementary flows;
- the Ecoinvent version 3.2 database (all four system models), provided in spreadsheet form via the ecoinvent website;
- The Ecoinvent implementation of 700 life cycle impact assessment methods with respect to Ecoinvent elementary flows.

In all cases, only publicly available information was included in the catalog. The catalogs themselves, as well as the software tools which created them, are available for inspection and use by the public.[2]

3.1. Catalog format

The catalogs described in this article are text documents written in JavaScript Object Notation (JSON), a plain-text format that can be easily read by both humans and computers. A more in-depth description of the entity descriptions, catalog descriptions, and tools for accessing them is found in the Supplementary Materials.

Each catalog includes the following fields:

- dataSourceReference: a path or hyperlink describing where the data were drawn from;
- dataSourceType: a text string describing the type of data source;
- processes: a list of process instances;
- flows: a list of flow instances;
- quantities: a list of quantity instances.

Each entity instance has the following fields which establish its identity:

- entityType must be either 'process', 'flow', or 'quantity';
- entityId is an identifier for the entity which is unique within the catalog;
- origin gives a description of the source repository containing the original data set;
- externalId gives the identifier used in the origin repository to refer to the data set.

Aside from these, each entity contains a collection of key-value pairs (e.g. “key”: ”value”) that describe the entity's *semantic* content, such as its name, geographic scope, classification, and all other information. The different entity types include certain tags at minimum, shown in Table 2. A more extensive list is provided in the supplementary materials.

Flow entities contain a list of “characterizations,” which are quantities that can be used to measure the flow. One of these entries may be marked as the reference quantity for the flow using the “isReference” tag. These characterizations include any life cycle impact characterization factors which are defined for the flows.

Process entities contain a list of “exchanges,” which are flows and directions that are observed to belong to the process. One or more exchanges may be marked with the “isReference” tag to signifiy the process's reference exchanges.

Because each entity has the same minimal structure, it is easy to manipulate the catalog in software to search for semantic content. Although the contents of the entry are generated by the data set maintainer, the format of the catalog allows entities from multiple data providers to be represented side-by-side and compared easily, either visually or in software.

3.2. Linking to the Semantic Web

The entities described by the catalogs, and also many of their properties, correspond to classes defined by ontologies that have been published on the Web. Our catalog establishes links between entity types and these web ontologies, enabling the catalog contents to be automatically associated with externally defined classes. This linking is established based on the JSON-LD specification (Sporny et al., 2014), which provides a way to convert JSON files into linked data graphs. A JSON-LD document can be easily transformed into RDF or other linked data formats. We created a *context* file that specifies the entity types and relationships encoded in the catalog by referencing published ontologies.

Table 2
Required semantic content fields for different entity types.

Entity type:	Quantity	Flow	Process
Reference entity:	Unit	Quantity	Exchange (flow, direction)
Name	X	X	X
Comment	X	X	X
Compartment		X	
SpatialScope			X
TemporalScope			X

Our contribution builds on the existing data model published as part of the OpenLCA Framework for common LCA concepts such as *flows*, *processes*, and *exchanges* (Ciroth and Sroka, 2014). Our contribution extends the schema with a new *quantity* entity, which generalizes existing flow properties and LCIA factors. Quantities contain a “referenceUnit” which specifies the unit of measure to be used in interpreting the quantity. In the future, these units should be defined by reference to externally managed units of measure that have commonly agreed standard meanings. Semantic Web research has produced a powerful ontology governing quantities and units of measure, unit conversions, uncertainty, and other aspects of quantitative data, called QUDT for “Quantities, Units, Dimensions, and Types” (Hodgson and Keller, 2011). Under our proposed model, the measurement quantities used in LCA could be directly linked to this ontology. Units of measure corresponding to impact assessment metrics do not have a standard Semantic Web representation. It is a project for future work to establish these links.

We also extend the OpenLCA schema by providing explicit links to meaningful concepts found in other ontologies, including spatial and temporal scopes, data set provenance, and other documentary information. The context file is documented in the supplementary materials.

3.3. Aggregation queries on the collection

Placing all databases into a common semantic data format enables a user to easily answer questions about the collection that are challenging when each database is presented in its own format. In this section we demonstrate a few queries that expose interesting information about the content and coverage of the databases.

3.3.1. Reference flows

Many databases were found to focus on certain technology classifications. Table 3 shows the most common reference flows (by count) found in each database. In this table, the Ecoinvent unallocated processes are used. US LCI is omitted because all reference flows in that database were all unique (with a few unremarkable exceptions).

In inspecting the results we see that nearly 25% of Ecoinvent processes generate electricity as the reference product, while a similar proportion of the GaBi professional database primarily generates steam or thermal energy. Many GaBi processes also report inputs as the reference flow; these processes are concerned with recovery of used materials at end of life. It is noted that there may be duplication across the GaBi professional database and extensions.

[2] https://bkuczenski.github.io/lca-tools-datafiles.

Table 3
Most frequently encountered reference flows in the data sources. GaBi extensions exclude the Ecoinvent v2.2 database.

	Ecoinvent (undef'd)	ELCD	GaBi pro	GaBi extensions
Total Number of Reference Flows	14,158	503	3309	7457
Output: electricity, high voltage	2350			
Output: Thermal energy (MJ)			236	944
Output: Electricity		64	471	522
Output: Steam (MJ)			622	340
Output: electricity, low voltage	730			
Input: Housing technology			191	340
Output: electricity, medium voltage	423			
Output: heat, district or industrial, other than natural gas	402			
Output: Cargo			127	80
Output: heat, district or industrial, natural gas	141			
Output: heat, central or small-scale, other than natural gas	139			
None: None		3	10	107
Input: Aluminium scrap			59	60

3.3.2. *Geographic coverage*

The geographic scope of the databases can also be compared in aggregate easily using the catalogs. The results for the fifteen most commonly cited geographic specifiers are presented in Table 4.

This table reveals the challenges associated with using simple text-based queries to consider geographic information. For instance, it can be seen that Ecoinvent predominantly uses 'RER' to signify the European region, while GaBi databases preferentially use 'EU-27' and ELCD uses a mix of the two. Without spatial reasoning it is challenging to understand the relationships among these signifiers.

The influence of research groups associated with Ecoinvent can also be seen: Switzerland is overrepresented in comparison to the rest of Europe, and Quebec ('CA-QC') is overrepresented in comparison to the rest of Canada (the "Canada without Quebec" region in the Ecoinvent database is used by only 3 processes).

4. Semantic applications of the catalogs

A central concern to much ongoing work in LCA data management is *interoperability*, which can be succinctly defined as "data exchange without significant information loss". In a way, data *providers* are already "interoperable" if they provide syntactic interoperability, i.e. if they provide information in a standardized format for data users to parse and interpret. The ILCD and Ecospold formats provide excellent examples of this form of interoperability, and interpretation of those formats was used to generate the catalogs above.

The challenge lies in making different data sets easily *interpretable* by users, which is where semantic tools are of value. Before computation of LCA results is even considered, practitioners must grapple with the much more elementary questions of *locating and validating scope-relevant data* for both the inventory and impact phases. A semantic catalog addresses this need by enabling data users to review the contents and coverage of different data sources before accessing the data sets themselves, and subsequently supporting the interpretation of terms encountered in the data sets.

Table 4
Geographic coverage of various databases.

	Ecoinvent (undef'd)	ELCD	GaBi pro	GaBi extensions	US LCI
All	13,307	503	3319	7457	701
GLO	6218	25	338	446	15
DE	168	19	314	2131	
US	92		137	1179	16
RNA	13		19	649	667
CH	1260	10	33	44	
RER	1136	75	84	14	3
EU-27		96	869	296	
CA-QC	346				
IN	60		55	187	
IT	73	11	52	149	
BR	66		59	153	
NL	76	10	65	117	
CN	65		99	99	
FR	94	10	50	106	
GB	70	10	62	97	

Furthermore, these semantic catalogs can aid the process of exploring how different databases use different terms and classifications to describe the similar processes and flows. They can be used as a platform to create and test better automated techniques to address interoperability and interpretability problems.

4.1. *Term Co-occurrence*

One important way meaning is encoded in a document is through the co-occurrence of important terms (Buzydlowski et al., 2002; Matsuo and Ishizuka, 2004). A data user searching for keywords in a data repository can gain useful information about the available data by observing what terms occur in tandem with the search term. However, simple search interfaces exposed by data providers are often inadequate to perform analysis based on term co-occurrence. When using the semantic catalog, a data user can compare the prevalence of keywords across multiple data sources.

For example, consider a data user interested in the modeling of vehicle emissions in the European context using the Ecoinvent and GaBi professional databases. He/she may know that European emissions directives are referred to by a sequence of standards known as "Euro 1", "Euro 2," and so on. Using the search interfaces provided on the GaBi and Ecoinvent websites, the user may have varying degrees of success locating documents and datasets that refer to these standards, but will gain limited intelligence about the overall representation of the standards in the databases. The search outcomes will vary from one site to the other, as will support for including wildcards or more complex search terms.

Using the catalogs, however, it is a simple matter to search with regular expressions, and to extract tags which appear frequently together with a given search term. Table 5 reports the results of a term-frequency query for the regular expression 'euro.?[0–9]' in the GaBi professional and Ecoinvent undefined databases respectively. The expression will match "Euro 4", "euro-5", "EURO1" and other similar terms equally. The results reveal interesting information about the contents of each database. For instance: Ecoinvent can be seen to model both passenger and freight emissions, while GaBi can be observed to model both rural and urban emissions. A data user can use the results to "drill down" into more

Table 5
Most frequently encountered tags with 'euro.?[0–9]'.

Ecoinvent term (num. processes)	GaBi term (num. processes)
SpatialScope = GLO (190)	Classifications = Processes (81)
Name = transport (173)	Classifications = Truck (79)
Name = freight (160)	Classifications = Road (76)
IsicClass = Freight transport by road (160)	Comment = driving share: HBEFA 3.1 (76)
IsicClass = Other passenger land transport (95)	SpatialScope = GLO (76)
Name = passenger car (95)	Comment = status January 2010– input parameter: Distance [km] (76)
TechnologyLevel = Current (91)	Comment = payload [t] (76)
TechnologyLevel = Undefined (82)	Comment = driving share motorway (76)
Name = market for transport (82)	Classifications = Transport (76)
SpatialScope = RER (65)	Comment = utilisation [–] (76)
Name = lorry with refrigeration machine (64)	Comment = rural (76)
Comment = internal combustion engine (61)	Comment = urban–average sulphur content: EU = 10 ppm (68)
Name = EURO4 (40)	Comment = sulphur content diesel [ppm] (58)
Name = EURO3 (40)	Name = Truck (50)
Name = EURO5 (40)	Name = Truck-trailer (20)
Name = R134a refrigerant (40)	Comment = –source emissions (12)
Name = EURO 5 (33)	Comment = non tampered (12)
Name = 3.5–7.5 ton (32)	Comment = –average emission values Euro 6 SCR (12)
Name = EURO6 (32)	Comment = –average emission values Euro 4 (12)
Name = carbon dioxide (32)	Comment = –average emission values Euro 3 (12)
Name = 7.5–16 ton (32)	Comment = –average emission values Euro 1 (12)
Name = EURO 4 (31)	Comment = –average emission values Euro 2 (12)
Name = EURO 3 (31)	Comment = –average emission values Euro 5 SCR (11)
TechnologyLevel = Modern (30)	

focused queries, or may combine multiple search terms to achieve more precise search results. Code to reproduce the search results is provided with the catalogs. Other information about data management is also exposed: much of the semantic content in the Ecoinvent database is expressed in process names, whereas the semantic content in GaBi is mainly stored in the comment field. This information can also be used to inform future searches.

4.2. Text similarity

A persistent challenge in using data from multiple databases is finding correspondences between similar or equivalent entities that are described differently. When matching flows from different databases, certain metadata fields, such as CAS number or formula, can be used to find matches automatically among a subset of flows (e.g. Ingwersen et al., 2015b). However, processes lack such generalized reference symbols, so it is often necessary to use unstructured text fields, looking for text descriptions that have similar semantic content.

There are several approaches that can help estimate the similarity of two different pieces of text. The simplest is to compute metrics that yield numerical scores based on the actual words in the process name. One such measure is the Jaccard Index (Nentwig et al., 2015) which is defined as the number of words common to both text strings, divided by the number of unique words. This reports the percentage of words in common for two process descriptions. Table 6 shows the results of this using the flow name of '*Roundwood, softwood, average, at forest road, NE-NC*' from the USLCI database. For the scores, we have subtracted the Jaccard Index from one so that lower scores are better, and a score of one means that there are no words in common.

What we see with this is that 'Road (average)' has the highest score since both the words 'road' and 'average' appear in the USLCI flow name, although they are only of minor importance in the meaning. The next eight results all have the same score since they all have only one word in common. The last result with a score of 1 indicates that there are no more flow names in the GaBi database that have words in common.

Table 6
Top 10 Jaccard Index scores for GaBi flow names compared to USLCI flow name '*Roundwood, softwood, average, at forest road, NE-NC*'. Lower scores indicate a better match. A score of 1 is the worst score, indicating no words in common.

GaBi flow name	Jaccard index score
Road (average)	0.714
Federal road	0.875
Land road	0.875
Municipal road	0.875
County road	0.875
Industrial road	0.875
Cement (average)	0.875
Softwood plywood	0.875
Softwood lumber	0.875
Crude oil, at consumer Ireland	1.0

Table 7
Top 10 Word2Vec scores for GaBi flow names compared to USLCI flow name '*Roundwood, softwood, average, at forest road, NE-NC*'. Lower scores indicate closer matches.

GaBi flow name	Word2Vec score
Timber cedar (12% moisture; 10.7% H_2O content) (m^3)	4.380
Timber spruce (12% moisture; 10.7% H_2O content)	4.380
Timber (12% moisture; 10.7% H_2O content)	4.380
Timber pine (65% moisture; 40% H_2O content)	4.525
Timber spruce (65% moisture; 40% H_2O content)	4.528
Road (average)	4.534
Laminated veneer lumber (LVL)	4.641
Wood pellets (5.8% H_2O content)	4.698
Waste incineration of untreated wood (10.7% H_2O content)	4.721
Solid construction timber (15% moisture)	4.738

Out of these results, only the ninth entry 'Softwood lumber' could be considered meaningfully similar. The Jaccard index, though simple to compute, does not perform well for small test strings with diverse vocabularies. This can be remedied somewhat by using query expansion (Voorhees, 1994) to augment the words in process names with synonyms. Online resources such as WordNet (Miller, 1995) can be used to locate synonyms that could improve searches across databases, although there is a risk that these could be applied too liberally and out of context, leading to erroneous matches. A further concern is maintaining the list of

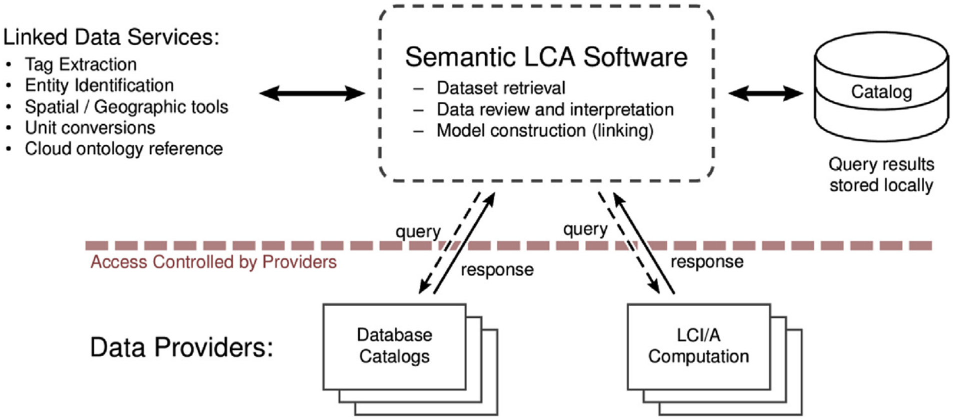


Fig. 2. A model for a semantically enriched LCA software system. Multiple data providers implement the above interfaces for data they own, controlling access if desired to licensed users. The user's software would build a local catalog of data resources across multiple providers and use semantic tools to help compare them.

synonyms.

Another approach is to use techniques that can somehow capture elements of the meaning or context of words. One promising approach which has emerged recently is Word2Vec (Mikolov et al., 2013). This machine learning algorithm generates vector representations of words via a neural network that tries to predict, given a specific word, what other words are likely to appear next in a text. Words whose vectors are close to each other are often synonyms or may be similar types of things (i.e. countries in the same region of Europe, etc). An advantage of this technique is that it is not necessary to maintain a list of synonyms because the similarity of words emerges automatically, given a large enough set of text to train the algorithm.

To use Word2Vec to evaluate the similarity of sequences of words such as process and flow descriptions, we employ a technique proposed by Kusner et al. (2015). For every word in one text string, we find the word in the second string that has the minimum distance to it. We then take the sum of these minimum distances for all the words in the first text string, and this is used as a measure of how close the first text string is to the second text string. The results of this are shown in Table 7 where we use the same flow name from the USLCI database mentioned above.

The Word2Vec approach of Kusner et al. (2015) is able to locate several terms that relate to wood: timber, cedar, spruce, pine, wood, veneer, lumber, and pellets, even though the processes have few or no words in common. What's further interesting is that many of the top results include species of trees that are indeed softwoods. A disadvantage of word2vec is its significant computational requirements. In practice, a suite of complementary techniques would probably be necessary to achieve the best results.

5. Discussion – toward semantic LCA software

Most LCA software systems in use today operate under a principle of stand-alone computation: an LCA researcher must procure a software system from a technology provider, and a complete, monolithic inventory database from another (possibly the same) provider. Impact assessment methods, although they are developed independently of LCA software systems, must be re-implemented by every software maker. The work presented in this paper suggests a novel approach, shown in schematic form in Fig. 2. In this concept, rather than providing a stand-alone database, data providers present a catalog interface to data users. Users interact with the catalogs by making queries to a semantic software system that interprets the user's requests and routes the queries to data providers that can answer them.

The query responses can be enriched through integration with linked data services that are in development or already exist. Some examples of how semantic data services can aid in the interpretation of life cycle data are provided in this paper, but there are also other opportunities. Unit conversion and quantity interpretation could be harmonized and simplified by making recourse to the QUDT ontology mentioned above, which is already mature. Tools for spatial reasoning have also been developed that could identify when one geographic region is contained within another, or could estimate the transport distances required for the outputs of one process to be made inputs to another. In the example in Table 1, spatial reasoning is required to determine that "RNA" is contained within "RoW" and not "RER".

Linked data services can also provide users with interpretive support by connecting query results to online semantic information resources. An example of a tool that does this already is DBpedia Spotlight (Mendes et al., 2011). This provides a web service where users can submit text and retrieve a list of entities extracted from the text (i.e. "diesel," "petroleum refining," etc) which are linked to their corresponding Wikipedia articles. Further enhancements could be provided by linking query results based on standardized product or industrial classification codes.

The use of catalog interfaces can transform users' interaction with private data. When a data user's interests are strictly qualitative, he/she will have access to rich semantic information describing the contents of available repositories. When quantitative information, such as exchange values or LCIA scores, are needed, the data providers can regulate access to this information by enforcing licensing requirements, thereby ensuring the value of data providers' investments in knowledge curation. Finally, the separation of semantic from quantitative content could improve metadata curation by allowing data users to revise and update linked data.

In practice, many of the challenges faced by LCA practitioners do not involve quantitative aspects of data: techniques for evaluating parametric sensitivity, performing Monte Carlo analysis, and propagating uncertainty estimates are well supported by existing software tools. In addition, data formatting and conversion tools are increasingly available (e.g. Ciroth, 2007). In contrast, variations in the structure and semantics of LCA data sharply constrain scientific progress in LCA in a much more fundamental aspect, by limiting the ability of users to discover and interpret available data. A semantic catalog of resources, publicly available and in a format suitable for extension and reuse, is a first step in empowering data users with much more powerful tools for reasoning and interpretation of LCA information.

Acknowledgments

The authors wish to thank the participants of the Santa Barbara 2015 GeoVoCamp for helping to generate the original ontology design patterns that inspired this work. The attendee list, as well as the products of the workshop, can be found at http://vocamp.org/wiki/GeoVoCampSB2015. We also thank the two anonymous reviewers for critical feedback which led us to significantly improve our work. The work was partially supported by NSF awards CCF-1442966 (Kuczenski) and ICER-1440202 (Janowicz). The funding agencies had no involvement in designing the research or preparing the article.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclepro.2016.07.216.

References

Ayres, R.U., 1995. Life cycle analysis: a critique. Resour. Conserv. Recycl. 14 (3–4), 199–223.

Baitz, M., Albrecht, S., Brauner, E., Broadbent, C., Castellan, G., Conrath, P., Fava, J., Finkbeiner, M., Fischer, M., Fullana i Palmer, P., Krinke, S., Leroy, C., Loebel, O., McKeown, P., Mersiowsky, I., Möginger, B., Pfaadt, M., Rebitzer, G., Rother, E., Ruhland, K., Schanssema, A., Tikana, L., 2013. LCA's theory and practice: like ebony and ivory living in perfect harmony? Int. J. Life Cycle Assess. 18 (1), 5–13.

Bertin, B., Scuturici, V.-M., Pinon, J.-M., Risler, E., 2012. CarbonDB: a semantic life cycle inventory database. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM'12. ACM, New York, NY, USA, pp. 2683–2685.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data – the story so far. Int. J. Semant. Web Inf. Syst. 5 (3), 1–22.

Buzydlowski, J.W., White, H.D., Lin, X., 2002. Term Co-occurrence Analysis as an Interface for Digital Libraries. Visual Interfaces to Digital Libraries, Springer Science + Business Media, pp. 133–144.

Ciroth, A., Sroka, M., 2014. OpenLCA Schema. https://github.com/GreenDelta/olca-schema.

Ciroth, A., 2007. ICT for environment in life cycle applications openLCA—A new open source software for life cycle assessment. Int. J. Life Cycle Assess. 12 (4), 209–210.

Cucurachi, S., Suh, S., 2015. A moonshot for sustainability assessment. Environ. Sci. Technol. 49 (16), 9497–9498. PMID: 26230673.

Davis, C.B., Nikolic, I., Dijkema, G.P.J., 2010. Industrial ecology 2.0. J. Ind. Ecol. 14 (5), 707–726.

Epimorphics, Ltd, 2010. ECO: the Earthster Core Ontology. http://www.epimorphics.com/web/projects/ECO.

Gangemi, A., 2005. Ontology design patterns for semantic web content. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (Eds.), The Semantic Web – ISWC 2005. Lecture Notes in Computer Science, vol. 3729. Springer Berlin Heidelberg, pp. 262–276.

Grüninger, M., Fox, M.S., 1995. Methodology for the design and evaluation of ontologies. In: The 1995 International Joint Conference on Artificial Intelligence (IJCAI-95).

Hák, T., Janoušková, S., Moldan, B., 2016. Sustainable Development Goals: a need for relevant indicators. Ecol. Indic. 60, 565–573.

Herrmann, I.T., Moltesen, A., 2015. Does it matter which life cycle assessment (LCA) tool you choose?—A comparative assessment of SimaPro and GaBi. J. Clean. Prod. 86, 163–169.

Hischier, R., Baitz, M., Bretz, R., Frischknecht, R., Jungbluth, N., Marheineke, T., McKeown, P., Oele, M., Osset, P., Renner, I., Skone, T., Wessman, H., de Beaufort, A.S.H., 2001. Guidelines for consistent reporting of exchanges/to nature within life cycle inventories (LCI). Int. J. Life Cycle Assess. 6 (4), 192–198.

Hitzler, P., Janowicz, K., Krisnadhi, A., 2015 (October). Ontology modeling with domain experts: the GeoVoCamp experience. In: Proceedings of the Diversity++ Workshop Co-located with the 14th International Semantic Web Conference (ISWC 2015).

Hodgson, R., Keller, P.J., 2011. QUDT-quantities, Units, Dimensions and Data Types in OWL and XML. http://www.qudt.org.

Ingwersen, W.W., Hawkins, T.R., Transue, T.R., Meyer, D.E., Moore, G., Kahn, E., Arbuckle, P., Paulsen, H., Norris, G.A., 2015a. A new data architecture for advancing life cycle assessment. Int. J. Life Cycle Assess. 20 (4), 520–526.

Ingwersen, W.W., Transue, T., Howard, T., Fowler, C., Meyer, D.E., Kahn, E., Arbuckle, P., 2015b. The LCA Harmonization Tool. LCA XV, Vancouver, CA.

Ingwersen, W.W., 2015. Test of US federal life cycle inventory data interoperability. J. Clean. Prod. 101, 118–121.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman II, C., 2014. Five stars of Linked Data vocabulary use. Semant. Web 5 (3), 173–176.

Janowicz, K., Krisnadhi, A., Hu, Y., Suh, S., Weidema, B., Rivela, B., Tivander, J., Meyer, D.E., Berg-Cross, G., Hitzler, P., Ingwersen, W., Kuczenski, B., Vardeman, C., Ju, Y., 2015 (October). A minimal ontology pattern for life cycle assessment data. In: Proceedings of 6th Workshop on Ontology and Semantic Web Patterns (WOP2015) Co-located with the 14th International Semantic Web Conference (ISWC 2015).

Kraines, S., Batres, R., Koyama, M., Wallace, D., Komiyama, H., 2005. Internet-based integrated environmental assessment using ontologies to share computational models. J. Ind. Ecol. 9 (3), 31–50.

Kraines, S., Batres, R., Kemper, B., Koyama, M., Wolowski, V., 2006. Internet-based integrated environmental assessment, Part II: semantic searching based on ontologies and agent systems for knowledge discovery. J. Ind. Ecol. 10 (4), 37–60.

Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q., 2015. From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 957–966.

Lenzen, M., Geschke, A., Wiedmann, T., Lane, J., Anderson, N., Baynes, T., Boland, J., Daniels, P., Dey, C., Fry, J., et al., 2014. Compiling and using input-output frameworks through collaborative virtual laboratories. Sci. Total Environ. 485–486, 241–251.

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends Ecol. Evol. 23, 159–168.

Matsuo, Y., Ishizuka, M., 2004. Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools 13, 157–169.

Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C., 2011. DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119.

Mila i Canals, L., Asselin, A.-C., Cruypenninck, H., Liu, S., Ingwersen, W., Macdonald, T., 2015 (May). Towards a global network of interoperable LCA databases. In: Proceedings of SETAC Europe 25th Annual Meeting (SETAC Europe 2015).

Miller, G.A., 1995. WordNet: a lexical database for English. Commun. ACM 38 (11), 39–41.

Muñoz, E., Capón-García, E., Laínez, J.M., Espuña, A., Puigjaner, L., 2013. Considering environmental assessment in an ontological framework for enterprise sustainability. J. Clean. Prod. 47, 149–164 (Cleaner Production: initiatives and challenges for a sustainable world {CP} Initiatives & Challenges).

Nentwig, M., Hartung, M., Ngonga, A., Rahm, E., 2015. A Survey of Current Link Discovery Frameworks. Semantic Web, pp. 1–18. Preprint.

Owens, J.W., 1997. Life-cycle assessment: constraints on moving from inventory to impact assessment. J. Ind. Ecol. 1 (1), 37–49.

Pauliuk, S., Majeau-Bettez, G., Müller, D.B., 2015. A general system structure and accounting framework for socioeconomic metabolism. J. Ind. Ecol. 19 (5), 728–741. http://dx.doi.org/10.1111/jiec.12306.

Pauliuk, S., Majeau-Bettez, G., Müller, D.B., Hertwich, E.G., 2015a. Toward a practical ontology for socioeconomic metabolism (online publication). J. Ind. Ecol. Adv.. http://dx.doi.org/10.1111/jiec.12386.

Sayan, B., 2011. The Contribution of Open Frameworks to Life Cycle Assessment. Master of Environmental Studies. University of Waterloo, Waterloo, Ontario, Canada.

Sonnemann, G., Vigon, B., Broadbent, C., Curran, M.A., Finkbeiner, M., Frischknecht, R., Inaba, A., Schanssema, A., Stevenson, M., Ugaya, C.M.L., Wang, H., Wolf, M.-A., Valdivia, S., 2011. Process on "global guidance for LCA databases". Int. J. Life Cycle Assess. 16 (1), 95–97.

Speck, R., Selke, S., Auras, R., Fitzsimmons, J., 2015. Life cycle assessment software: selection can impact results. J. Ind. Ecol. 20 (1), 18–28.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., Lindström, N., 2014. JSON-LD 1.0 – a JSON-based Serialization for Linked Data. https://www.w3.org/TR/json-ld/.

Takhom, A., Suntisrivaraporn, B., Supnithi, T., 2013. Ontology-enhanced life cycle assessment: a case study of application in oil refinery. In: The Second Asian Conference on Information Systems (ACIS), Phuket, Thailand.

United Nations, 2014. The Road to Dignity by 2030: Ending Poverty, Transforming All Lives and Protecting the Planet. Synthesis Report of the Secretary-General On the Post-2015 Agenda. Tech. rept., The United Nations.

United Nations, 2015. The Millennium Development Goals Report 2015. http://www.un.org/millenniumgoals/.

Voorhees, E.M., 1994. Query Expansion using Lexical-Semantic Relations. SIGIR'94. Springer, London.

Yan, B., Hu, Y., Kuczenski, B., Janowicz, K., Ballatore, A., Krisnadhi, A.A., Hitzler, P., Suh, S., Ingwersen, W., 2015 (October). An ontology for specifying spatiotemporal scopes in life cycle assessment. In: Proceedings of the Diversity++ Workshop Co-located with the 14th International Semantic Web Conference (ISWC 2015).

Zhang, Y., Luo, X., Buis, J.J., Sutherland, J.W., 2015. LCA-oriented semantic representation for the product life cycle. J. Clean. Prod. 86, 146–162.