Creating LCA Data Exchange Networks

Oliver Kusche\textsuperscript{1}, Clemens Düpmeier\textsuperscript{1}, Marco Recchioni\textsuperscript{2}, Fabrice Mathieux\textsuperscript{2}

Abstract

In the Integrated Product Policy Communication of 2003, the European Commission recognised Life Cycle Assessment (LCA) as "the best framework for assessing the potential environmental impacts of products". It also identified "the need to improve data availability and quality worldwide by internationally cooperating on LCA data and methods". To address this situation and to foster standardization, the ILCD (International Life Cycle Data) format was developed by the European Commission's Joint Research Centre (DG JRC) and implemented by the Institute of Applied Computer Science (IAI) of the Karlsruhe Institute of Technology (KIT). IAI also implemented an LCA database application for storing ILCD datasets by the name of soda4LCA (service oriented database application for LCA), which was published under an Open Source license. Subsequently, the idea of establishing a global, decentralized data network for high quality LCA datasets based on the ILCD format called the ILCD Data Network has been proposed and first data network features were implemented in soda4LCA. This paper summarizes requirements for implementing LCA data networks and presents the already implemented network features as well as an outlook to future enhancements of soda4LCA. It also presents how these features are instrumented by the ILCD Data Network.

1. Introduction

In the Integrated Product Policy Communication of 2003, the European Commission recognised Life Cycle Assessment (LCA) as "the best framework for assessing the potential environmental impacts of products". It also identified "the need to improve data availability and quality worldwide by internationally cooperating on LCA data and methods". In 2008, life cycle approaches were further strengthened in EU policies through the Sustainable Production and Consumption/Sustainable Industry Policy Action Plan Communications that encompasses various policies (e.g. Eco-design for Energy-related Products Directive). The life cycle approach is also a key approach of the 2011 Communication on "A resource-efficient Europe – Flagship initiative under the Europe 2020 Strategy".

While nowadays many LCI and LCIA data are available, via software tools as well as on the Internet, practitioners still experience difficulties searching for and finding needed datasets. Often, the original dataset provider may be difficult to identify and the quality of the dataset difficult to judge. Often it's also difficult to find meta-information and accompanying background documents that give sufficient information about how to use the dataset or its underlying process model.

To address this situation and to foster standardization and collaboration through joint data usage, the ILCD (International Life Cycle Data) was developed by DG JRC and the idea of establishing a global, decentralized data network for high quality LCA datasets has been proposed and initiated by the JRC as the ILCD Data Network (Wolf/Chomkhamsri/Düpmeier/Kusche/Pant/Pennington 2010).

\footnote{1}Karlsruhe Institute of Technology (KIT), Institute for Applied Computer Science (IAI), Herrmann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

\footnote{2}European Commission, Joint Research Centre (JRC) Institute for Environment and Sustainability (IES), via Enrico Fermi 2749, T.P. 27, JRC, 21020 Ispra (Va), Italy
The objective of data networking in this context is a harmonized data system, where reference data (as described above) and common base process data can be provided over the network to all parties of the network. While each network partner maintains control about his private working data, the operator of one node can decide to publish some or all of his datasets – like the common background data - as a dataset package available over the network for access by others. As shown later, the soda4LCA software implementing the ILCD network will then synchronize such shared packages to all nodes which have interest to provide these datasets as well to their users.

But there are also other use cases for LCA data networks and each of them has their own requirements as discussed in the next chapter.

2. Requirements for data networking

In this section, a few other use cases and application scenarios are discussed and corresponding requirements derived.

Commercial Usage

While there is a great variety of individual licensing models for LCA data, in principle, a distinction can be drawn between data that is published free of charge and, on the other hand, data that is provided for a fee. In a data network, there should be a possibility to handle both types of datasets.

To address this situation, a mechanism is needed that allows for limiting access for certain users (e.g. the anonymous user accessing a public database) to only the metadata of a process dataset, enabling them to inspect the metadata and then for instance turn to the dataset vendor to obtain proper permissions to access the full dataset.

With providing an option to publish only metadata and charge a fee for the full dataset, commercial scenarios can be addressed properly.

Academic research

In academic research projects, usually only a few datasets are generated, which rely on freely or commercially available base data (like the ecoinvent or ELCD databases) as a foundation. Researchers may want to share a working platform to coordinate work on datasets in research projects. If no such platform is available, conflicts may arise at the time when results are combined or merged if each partner uses different (and possibly incompatible) versions of base data.

Enterprise

In an enterprise environment, possible applications include internal distribution of both base data and internal datasets among subsidiaries or working groups as well as retrieval, acceptance or rejection of unit process datasets coming from suppliers. Furthermore, aggregated data may need to be passed to industry associations. Published data should only include highly aggregated data while hiding process details.

Industry association

Industry associations usually collect data from their members which are aggregated either by the association or the member (see above) before it is published. Such organizations may want to publish data or feed data into other networks from their own servers in order to be able to deliver timely updates.
Institutional database

There may be institutions (such as, for example, the European Commission) that wish to provide a large collection of multi-sectorial data that should be available free of charge. Since it is intended to be used for policy making, it is important that this data is quality-assured. Institutional databases often combine data from different sources, such as industrial associations. It’s easier for such institutions if the delivery of data and the merge process can be supported by automatic mechanism provided by a data network.

From these scenarios, the following aspects can be derived as key requirements for data networking:

- **Collaboration** - Dataset generation and publication sometimes take place in a highly decentralized environment, across physical locations and organizations. For successful collaboration on data, jointly used data must be easily accessible. As a distinction is drawn between private working data and data for publication, a flexible model for access is needed.

- **Dataset stability and Dataset quality** - One of the main issues with joint data usage is the distribution of data and especially updates, since compatibility between versions must be ensured. Datasets must be stable once published, otherwise when referenced in studies; changes of the dataset may render study results invalid. This defines needs for a version concept in which one version of a dataset can be accessed from outside for studies while there is already work on another version of this dataset.

Because dataset quality is crucial for applications, there should be a clear distinction between datasets which have release quality and are released (these are the ones which can be used by others) and datasets which doesn’t yet have release quality or being worked on.

- **Quality control and harmonization** – In larger data networks managed by institutions there are needs to formalize the quality control of datasets created across a network by different parties but then compiled into one harmonized database through reviews done by a review panel and other organization means. The data network software should provide adequate functionalities to support this use case.

The above requirements led to several enhancements of the storage model for LCA datasets in soda4LCA, the database application developed by the IAI for DG JRC, to implement the ILCD data network. In the next chapter these enhancement will be described first before the data network concepts implemented in soda4LCA will be discussed in the following chapter.

3. **The soda4LCA database software**

The soda4LCA (service oriented database application for LCA) database software that has been previously presented (Düpmeier/Kusche 2011) has been developed at the IAI as a flexible database application for LCA data in the ILCD format.

The software has a web-based user interface that can be accessed with any web browser as well as a REST service interface. The latter can be used by other software applications or other soda4LCA instances to access a certain soda4LCA instance across the Internet as illustrated by Figure 1. Soda4LCA can store ILCD-formatted process (unit and aggregated processes, results), LCIA method, flow, flow property, unit group, contact and source datasets and referenced documents in the database. As shown in Figure 1, soda4LCA includes functionalities for import and export and search and retrieval of datasets. An access right management system allows administrators to define which users have what kind of access (read, read only metadata, import, overwrite) to datasets. A further dataset administration module allows the management of datasets in a soda4LCA database instance.
Central concepts for data management and also for the data networking functionalities in soda4LCA are the concepts of dataset versions, their versioning and release management and the concept of data stocks. These concepts are described in the following sections.

3.1 Data Management

In this section, the concepts implemented in soda4LCA that are fundamental to dataset management are described.

3.1.1 Working copies vs. released datasets

When datasets are residing in remote locations and are referenced via links over the internet, data consistency and stability have to be ensured. Once a dataset is published over a data network, it must not ever change; otherwise any studies relying on that dataset may be rendered invalid, depending on the kind and extent of changes. On the other hand, datasets do need to be updated, for example to fix errors or include new findings or more accurate results.

To overcome this dilemma, soda4LCA supports the storage of different versions of a logical dataset and each version can be given a status that may be either RELEASED or UNRELEASED. When a dataset is first imported into the database or copied from another existing dataset, it is regarded as a working copy and thus carries the status UNRELEASED. Being a work in progress, it can be altered and updated and is local to its node. When a dataset is to be published, its state can be set to RELEASED. This means that this particular version of the dataset is frozen and now immutable, i.e. cannot be altered anymore. Datasets carrying this state may then be distributed across the network.

To update a dataset that carries the status RELEASED, a new version has to be made, whose status will automatically be set to UNRELEASED, and stored with the incremented version number.
3.1.2 Data stocks

The concept of data stocks is an important concept for logically grouping datasets together. A data stock is a set of datasets that belong together in some context. It can comprise an arbitrary number of datasets.

Every dataset in a soda4LCA instance is associated to exactly one root data stock. As the dataset is changed, all working versions will remain associated with that root data stock.

Arbitrary datasets that carry the RELEASED status may be grouped together to form a release data stock. As a dataset with the status RELEASED is immutable, a dataset from a release data stock is guaranteed to not ever change.

Moreover, the data stock itself carries a state of RELEASED or UNRELEASED with the same purpose – once set to RELEASED status, datasets cannot be removed from a data stock anymore (while more datasets may still be added).

A release data stock may only contain datasets with distinct UUIDs, i.e. cannot contain multiple versions of the same dataset. Release data stocks may be marked as exportable to other network nodes in which case they and their datasets will be synchronized between network nodes (see the next chapter).

A released data stock may include datasets from arbitrary nodes. The necessary condition for this is that each dataset is part of a released and exported data stock of these nodes.

Further information about data stocks and version management can be found in another paper presented at the EnviroInfo 2012 conference (Düpmeier/Greceanu/Kusche/Schmitt 2012). The next chapter will describe how the data stocks concept are related to the data network concepts of soda4LCA.

4. Concepts for building data networks with soda4LCA

A database node with soda4LCA can be operated independently as well as joined with other nodes to form a network. Currently, there are two different modes how nodes can connect to each other: peer-to-peer networking and registries networking mode.

Peer-to-peer Networking Mode

In peer-to-peer networking mode (Figure 3), each node holds an individual list of other nodes, which is manually maintained by the node administrator. To add a foreign node to the list, its base service URL (and, optionally, credentials for authentication) is entered. The other node will then be contacted by the application and, if successful, its metadata is retrieved and the new node is added to the list of known nodes.

As each node only knows about the nodes on its list, the relationship is in principle unidirectional and it is neither transparent to a node which other nodes are connected to it nor which other nodes exist on the network. Node administrators control which other nodes can connect to their node through authorization.

Registry Networking Mode

In order to enable nodes to know about other nodes on the network as well as to maintain control about joining nodes, a separate registry facility that maintains a list of all nodes on the network was introduced where nodes may register (Figure 4). If a node is to be registered with the registry, a node registration request is send to the registry and then has to be approved by the registry administrator. Upon successful registration, the new node will be added to the list of registered nodes, which is propagated to all nodes on the network. A node may be registered in multiple registries.

Thus, after registration on a data network working in registry networking mode, the registering node has access to all nodes available on the data network and vice versa.
When a user chooses the option to perform a distributed dataset search on a data network setup with the current release version of soda4LCA, all nodes in his node list will be queried with the same query parameters. Each node, which is available at the time of the query, will return a result list of results matching the query. The results from all the nodes will then be cumulated and returned along with the results from the local node. This usage of a data network is far from optimal.

First, data managers on the different nodes don’t have control over which of their datasets will be accessible to others. Second, the distributed search for datasets depends on the network availability of the nodes. If a node is down while a search is performed results of this node cannot be returned. Also the performance of the search is not optimal if many nodes are involved because it depends on the performance characteristics of each node.

Because of this the network data access will be re-implemented in the current development version of soda4LCA using the already existing data stock concept described briefly in chapter 3.1.2. In the new upcoming version of soda4LCA release data stocks on one node can be marked as replicable and accessible by other nodes in a data network by data managers or node administrators. A notification will then be send to others nodes informing the node administrators that new data collections (stocks) are available on the data network. A node administrator of such a node can then import datasets from such a release stock from another node either by importing only some of the released datasets from this node into another data stock at their own node or by importing the whole release data stock as complete dataset bundle. In both situa-
tions the referenced datasets will then be replicated (i.e. copied) to the referencing node. This solves exist-
ing performance problems and makes the data network far more robust.

5. The ILCD Data Network as an Example Implementation

The International Reference Life Cycle Data System (ILCD) Data Network aims at providing consistent
and quality-assured LCI data with easy access via searches, filtering, and sorting. The datasets in the net-
work can come from any data developer/owner, e.g. industry, national LCA projects, research groups, and
consultants. The data is to be published by the developer or owner under their own terms and conditions,
e.g. free of charge, for a fee, via registration, etc. The development of datasets and their publication are
independently managed by the respective data developers.

As part of the ILCD Data Network, the datasets of the European Reference Life Cycle Database
(ELCD) - over 300 LCI datasets of European scope as well reference datasets like a reference elementary
flows, standard flow properties and associated units - are foreseen to be made available as one contribu-
tion from Europe. The ELCD provides LCI data for basic materials, energy carriers as well as transport
and end-of-life treatment services. This database includes a large number of official data from currently 16
EU-level industry associations.

5.1 Requirements

To establish a meaningful Data Network, as opposed to a library of methodologically and IT-wise incom-
patible datasets or a directory of databases, from the beginning the need exists to operate with a minimum
set of requirements. In order to be accepted in the ILCD Data Network, datasets should meet some data
quality requirements. Only datasets compliant with the "ILCD Data Network - entry-level requirements"
(JRC 2010) are part of the network. This level of requirements has been put in place to quickly increase
the access to properly documented LCI datasets, working with a common set of elementary flows etc. At
the same time are the datasets identified according to the quality-level they adhere. It is important to high-
light that all data in the network have to be in line with ISO 14040 and 14044, that is part of the entry-
level requirements. In the establishment phase of the ILCD Data Network, there is the need to meet the
ILCD Data Network entry-level requirements that have been published by JRC.

5.2 Features and Current Status

JRC has developed the ILCD registry application that enables a networking scheme as described in section
4 - Registry Networking Mode to allow implementing previously presented scenarios. The registry applica-
tion has been contributed for integration into soda4LCA, which is the technical platform for the net-
work’s nodes operation.

The registry functionality can be leveraged to manage a network of soda4LCA nodes by an administra-
tor entity. Node managers can apply to the registry for both node registration and datasets registration.

The registry functionality has been developed to allow the network administrator to have control on
nodes and datasets to be published: the registry administrator has the possibility to accept or reject node
registration and datasets submission, if minimum quality requirements are not met or if information is
lacking. The network search functionality within the registry therefore guarantees the user to download
quality ensured datasets. This functionality allows the implementation of previously described scenarios.

The ILCD Data Network soon to be launched will allow LCA and environmental footprint practitioners
to better link with developers of life cycle datasets when ensuring transparency concerning quality and
consistency for data.
6. **Summary and Outlook**

On the basis of different scenarios where networking of LCA data is necessary or deemed beneficial, relevant requirements for LCA data networks have been highlighted. Fundamental dataset management concepts as they are implemented in soda4LCA have been shown as well as the networking capabilities of the latest released soda4LCA version (1.3). Then an outlook was given on the re-designed data network feature of the new upcoming development version of soda4LCA and finally it was demonstrated how these features can be applied to fully implement JRC’s vision of the European ILCD Data Network.

While the main work on soda4LCA is now focused on implementing the complete model of data stocks and the networking features explained above there is also work foreseen for the next version of soda4LCA which will implement support for Environmental Product Declaration (EPD) datasets in soda4LCA.

The Authors would like to thank Marc-Andree Wolf for his significant contribution to the concept of the registry and to the development of the software implementation.

**Bibliography**


Wolf, M.A.; Chomkhamsri, K.; Düpmeier, C.; Kusche, O.; Pant, R.; Pennington, D. - International reference life cycle data system (ILCD) data network. Internat.Conf.on EcoBalance, Tokyo, J, November 9-12, 2010


Ciroth, A.; Srocka, M.; Düpmeier, C.; Kusche, O.; Schebek, L. - Interacting with the ILCD data network from an LCA software - the example of openLCA. Life Cycle Management Conf.2011 (LCM 2011), Berlin, August 28-31, 2011


Copyright 2012 Shaker Verlag Aachen, ISBN: 978-3-8440-1248-4