



Wir schaffen Wissen – heute für morgen

Paul Scherrer Institut

Chris Mutel

Can't we all get along?

The pain and promise of LCA data interchange

The problem



- Different software cultures
- Different expectations
- Different worldview
- Good people
- Limited resources
- System evolved

Star Trek image copyright someone really important but not sure who please don't sue me

2

The current data format and interchange situation is the result of years of incremental changes in a divided software marketplace. Everyone has tried to make things easy and correct for their customers and their software, but the end result, at least for data interchange, is a bit of a mess.

Ecospold 1

```
<flowData>
  <exchange number="1" infrastructureProcess="false" name="Dummy, agrochemicals, at plant"
    location="RNA" unit="kg" meanValue="0.52"
    generalComment="See pesticide use summary sheet.">
    <inputGroup>5</inputGroup>
  </exchange>
  <exchange number="2" infrastructureProcess="false"
    name="Diesel, combusted in industrial equipment" location="RNA" unit="L"
    meanValue="14.3">
    <inputGroup>5</inputGroup>
  </exchange>
```

3

For example, the ecospold 1 data format seems easy enough, but simply fails as a data interchange format (extract from 2014 US LCI). The ‘number’ field is used differently by different software, so is unreliable. Several necessary fields, such as categories, are sometimes simply omitted. XML is a strict format, but not strict in the right way - there is no guarantee that names are used consistently and without typos, and no guarantee that a process called “Diesel, ...” is actually available in this database export. Indeed, the “dummy” processes are explicitly not included.

Example: US LCI (2014)

```
In [16]: sp.statistics()
1204 datasets
31262 exchanges
10040 unlinked exchanges
  Type biosphere: 1154 unique unlinked exchanges
  Type production: 278 unique unlinked exchanges
  Type substitution: 1 unique unlinked exchanges
  Type technosphere: 243 unique unlinked exchanges
Out[16]: (1204, 31262, 10040)
```

- Missing (links to ecoinvent 2.2)
- Wrong names: “Electricity, at grid, Eastern US” -> “Electricity, at grid, Eastern US, 2000” (> 40)
- Wrong locations: “Electricity, at grid, US, 2000” US -> RNA
- Biosphere flow mismatches: “Carbon dioxide” must be fossil or non-fossil

4

A data exchange format must produce the same results on multiple software systems. But as the US LCI shows, the ecospold 1 format makes no guarantees that the processes in each file can actually be linked together. Screenshot is from [brightway2-io](https://brightway2-io.github.io/). See [importing the US LCI notebook](#).

SimaPro CSV

```
Process name  
[sulfonyl]urea-compounds, at regional storehouse/kg/CH
```

```
Materials/fuels  
Lubricating oil, at plant/RER U;kg;0,  
000514;Lognormal;1,22;0;0;"(1,5,2,3,1,3); "  
Building machine/RER/I U;p;0,000000134;Lognormal;3,  
05;0;0;"(1,5,2,3,1,3); "  
Diesel, at regional storage/RER U;kg;0,  
0234;Lognormal;1,22;0;0;"(1,5,1,3,1,3); "
```

So much sadness in the world...

5

“What can I say about that suit that hasn't already been said about Afghanistan?”

SimaPro CSV is not only not really CSV, but it is also not documented (see [a reverse engineering attempt](#)), changes frequently, and linking even within a single file is not guaranteed.

- Names, locations, and other metadata are combined instead of being stored separately
- There is no guarantee that product names are unique within a file, so you could have two processes which produce “foo”, and no way to know which one to link to if “foo” is an input
- SimaPro CSV is stored in the [Windows-1252](#) text encoding, and converts unknown characters to “?”, making data roundtrips impossible.

Return to sanity...

```
<flowData>
  <intermediateExchange id="f4208cd4-26ae-45f8-a5ac-5382c77c7196"
    unitId="487df68b-4994-4027-8fdc-a4dc298257b7" casNumber="9002-88-4" amount="-67500"
    intermediateExchangeId="7ee0ddc3-924d-4339-b9bf-7856af6a698f"
    activityLinkId="e13e6a26-ef81-491a-a5b2-e7174ccede41" productionVolumeAmount="202500">
    <name xml:lang="en">waste polyethylene</name>
    <unitName xml:lang="en">kg</unitName>
    <comment xml:lang="en">(4,5,1,3,4,5); Calculated, based on input</comment>
    <comment>EcoSpold01Location=CH</comment>
```

- Links use UUIDs
 - Can have same name, but different UUIDs!

But

- No guarantee of internal consistency
- Updating dataset links
- Specific to ecoinvent view of world

Ecospold 2 uses machine-generated unique identifiers, but there is still no guarantee of internal consistency, and is way to ecoinvent 3-world specific.

There is also no provision to indicate data updates; UUIDs do not change on data updates.

Ecoinvent 3.1 default

We can do the same thing with the default system model:

```
In [11]: ei3ldefault = SingleOutputEcospold2Importer(
          "/Users/cmutel/Documents/LCA Documents/Ecoinvent/3.1/default/datasets",
          "ecoinvent 3.1 default"
        )
ei3ldefault.apply_strategies()
ei3ldefault.statistics()

Extracting ecospold2 files:
0%          100%
[#####] | ETA[sec]: 0.000 | Item ID: ffed8e5b-8ecb-4
Total time elapsed: 104.858 sec

Title: Extracting ecospold2 files:
Started: 05/18/2015 11:15:39
Finished: 05/18/2015 11:17:24
Total time elapsed: 104.858 sec
CPU %: 87.300000
Memory %: 7.371628
Extracted 11329 datasets in 105.83 seconds
Applying strategy: remove_zero_amount_coproducts
Applying strategy: remove_zero_amount_inputs_with_no_activity
Applying strategy: es2_assign_only_product_with_amount_as_reference_product
Applying strategy: assign_single_product_as_activity
Applying strategy: create_composite_code
Applying strategy: drop_unspecified_subcategories
Applying strategy: link_biosphere_by_flow_uuid
Applying strategy: link_internal_technosphere_by_composite_code
Applying strategy: delete_exchanges_missing_activity
Applying strategy: delete_ghost_exchanges
116 exchanges couldn't be linked and were deleted. See the logfile for details:
/Users/cmutel/Library/Application Support/Brightway3/ecoinvent-import.99d620c13d8ce5ab521afdd7alc6ab75/logs/E
cospold2-import-error.bVBLjP.log
11329 datasets
806149 exchanges
0 unlinked exchanges

Out[11]: (11329, 806149, 0)
```

In ecoinvent 3.1 default, there are 116 exchanges which link to a product that is not produced by the indicated activity (production amount is zero). Each software system will invent its own way to handle this problem, leading to inconsistencies. Also note the number of strategies (data transforms) which are needed before linking across files is possible.

Brightway2-IO

Aluminum, primary, smelt, at plant			kilogram	Primary Meta	RNA		
Name	Amount	Database	Unit	Categories	Location	Type	Matched
Alumina, at plant	1930	US LCI	kilogram		RNA	technosphere	TRUE
Aluminum, primary, smelt, at plant	1000	US LCI	kilogram		RNA	production	TRUE
Ammonium, ion	0.00057	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Anode, at plant	455	US LCI	kilogram		RNA	technosphere	TRUE
BOD5, Biological Oxygen Demand	0.0091	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
CFCs and HCFCs, unspecified	0.12		kilogram	air	(unknown)	biosphere	FALSE
COD, Chemical Oxygen Demand	0.079	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Carbon dioxide, fossil	1520	biosphere3	kilogram	air	(unknown)	biosphere	TRUE
Carbon monoxide	66.9		kilogram	air	(unknown)	biosphere	FALSE
Carbonyl sulfide	1.12	biosphere3	kilogram	air	(unknown)	biosphere	TRUE
Chloride	0.00842	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Cyanide	0.0002	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Detergents, unspecified	0.00059		kilogram	water	(unknown)	biosphere	FALSE
Diesel, combusted in industrial boiler	1.836	US LCI	litre		RNA	technosphere	TRUE
Dissolved organic matter	0.013		kilogram	water	(unknown)	biosphere	FALSE
Dissolved solids	0.076	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Dummy_Disposal, solid waste, unspecified, to unspecified treatment	59.39	US LCI	kilogram		RNA	technosphere	TRUE
Electricity, aluminum smelting and ingot casting regions	15410	US LCI	kilowatt hour		RNA	technosphere	TRUE
Ethane, hexafluoro-, HFC-116	0.0228	biosphere3	kilogram	air	(unknown)	biosphere	TRUE
Fluoride	0.051	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Gasoline, combusted in equipment	0.28374	US LCI	litre		RNA	technosphere	TRUE
Hydrocarbons, unspecified	0.0000048	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Hydrogen cyanide	0.037		kilogram	air	(unknown)	biosphere	FALSE
Hydrogen fluoride	0.62	biosphere3	kilogram	air	(unknown)	biosphere	TRUE
Iron	0.0022		kilogram	water	(unknown)	biosphere	FALSE
Lead	0.0000046	biosphere3	kilogram	water	(unknown)	biosphere	TRUE
Liquefied petroleum gas, combusted in industrial boiler	5.007	US LCI	litre		RNA	technosphere	TRUE
Mercury	0.0000004	biosphere3	kilogram	water	(unknown)	biosphere	TRUE

Brightway2-io (link on slide 4) can export databases to spreadsheets, and indicate unlinked exchanges.

Problem 1: Naming things

- Need **one** correspondence list between elementary flows
 - Can be 1 to 1, N to 1, 1 to N
 - Don't have to wait for software/database developers
- LCIA method developers have to link to this list

9

We desperately need a master list of biosphere flows, and how to translate between the different lists used in ILCD, SimaPro, OpenLCA, ecoinvent, etc. [olca-converter](#) is great, but not enough.

Problem 2: Linking

- Links should be:
 - Clear
 - Verified
 - **No composite links** (e.g. name & location)
- Links **can't** use metadata
 - No guarantee of uniqueness
- Linking software should be open

10

Naming things is a fundamental problem in computer science, and a lot of smart people have thought deep thoughts about names. We don't have to reinvent the... well, anything.

Linking should be based on one machine-generated unique ID, and should be verified on database export. Linking software, especially for complicated system models, must be open so that different software systems can consistently import databases the same way.

Problem 3: Updating datasets

- Each revision gets a new unique id
 - Linking is to **this** unique id
- A database consists of process datasets **& upgrade path**
 - Upgrade path must indicate breaking changes

Joe Armstrong: [The mess we're in](#)

11

As far as I know, no software does this yet, but it is crucial in a world of distributed data generation and analysis.

Conclusions

- Problems come from:
 - Error-prone linking
 - Undocumented formats
 - No database-level checks or constraints
 - Ad hoc systems
- Solutions include:
 - More LCA software
 - Well documented migrations/upgrades
 - Single unique identifiers tied to specific versions
 - JSON-LD

12

[JSON-LD](#) is really great - not only is it a more modern format, but it also will lower the barrier of entry for writing LCA software, as JSON is the language of the web.

More information:

- Email: cmutel@gmail.com
- Website: brightwaylca.org
- Development blog: chris.mutel.org

