Chris Mutel :: Paul Scherrer Institut

inventories.io: A new format and network for inventory data

Ecobalance 2016, 4 October 2016

Handling LCA data is a pain

- Formats
- Names
- Linking

- Changes
- Updates
- Conflicts

Database versus userland

Photo credit: amenclinicphotos ac

Moving LCA data between different systems (or even managing data migrations and updates inside our existing systems) if a huge pain.

- There are small but significant implementation differences in common formats
- Names and other metadata are modified from software to software
- Linking a foreground exported from OpenLCA to a background exported from SimaPro, for example, is not simple. See also the Brightway2 IO notebooks
- There a few common and documented procedures to handle dataset changes or updates, especially when transferring data between computers

The current system works, at least to some degree, if you stay within a given software ecosystem. However, our future is one where many people contribute on many levels, including with small and large software libraries, and that requires true and pain-free data interchange.

SimaPro CSV

Linking

- Links to flows not activities
- Links to text strings
- No versioning
- No provenance

Sharing

- **Broken** ecoinvent
- Single files per export

```
Materials/fuels;;;;;;;;
Passenger car maintenance {GLO}| market for | Alloc Rec,
U;6.45161E-06;p;Lognormal;1.0287;;;unchanged;
```


SimaPro CSV has some good qualities - linking to products or flows instead of activity names, and the ability to include an entire dataset within a single file - but does not provide enough information for deterministic linking between datasets, nor any ability to mark where a dataset came from, or whether it was modified.

Ecospold 1

Linking

- Links to text strings
- No activity/flow distinction
- No resource locaters for external links
- No check for integrity

Sharing

- Single file
- No normalised form

```
<exchange number="414" category="chemicals" subCategory="
organics" localCategory="Chemikalien" localSubCategory="
Organisch" name="latex, at plant" location="RER" unit="kg"
uncertaintyType="1" meanValue="5.4" standardDeviation95="1.05"
generalComment="(1,1,1,1,1,1,4)" localName="Latex, ab Werk"
infrastructureProcess="false">
    <inputGroup>5</inputGroup>
</exchange>
```


Ecospold 1 is even worse. These formats were designed by well-intentioned people, and were very important in getting our community to where it is today, but do not address some specific modern data needs and technologies.

Ecospold 2

Linking

- Links to flows not activities (UUID for activity and flow)
- No resource locaters for external links

Sharing

- Single file
- No normalised form

```
<intermediateExchange id="
8d9cb33d-148f-416b-8d7d-12f1210cb625" unitId="
77ae64fa-7e74-4252-9c3b-889c1cd20bfc" amount="
0.171654366842251" intermediateExchangeId="
66c93e71-f32b-4591-901c-55395db5c132" activityLinkId="
f0f93629-29a7-4bb4-bd25-de63ab69658f"
productionVolumeAmount="66563000000">
  <name xml:lang="en">electricity, high voltage</name>
  <unitName xml:lang="en">kWh</unitName>
```


Ecospold 2 is strictly better than Ecospold 1, although it does add a lot of complexity. However, it still does not have a single UUID for an activity, and can’t link in a meaningful way to flows not present in the database export.

ILCD

Linking

- Links to flows not activities
- Links are versioned (activities as well), e.g. "`1.0.0`"
- Links are relative URIs, e.g. "`../flows/Electrical_power_…xml`"
  - Assumes same computer

Sharing

- **Many** files per activity
  - OK for entire databases, **death** for single activities
- Good document on use of UUIDs (and when they should change)

ILCD is quite good, especially for exporting large databases. However, it doesn't really work for sharing a single dataset, where we want to be able to point to a single file-like object in a data sharing system.
The ILCD document on managing UUIDs is quite good.

Requirements for data interchange

- Deterministic linking
  - Exact identification
  - Included to datasets not in export
- Change history
  - Format safe for DVCS (e.g. git)
  - Branching
- Data integrity
- Effortless

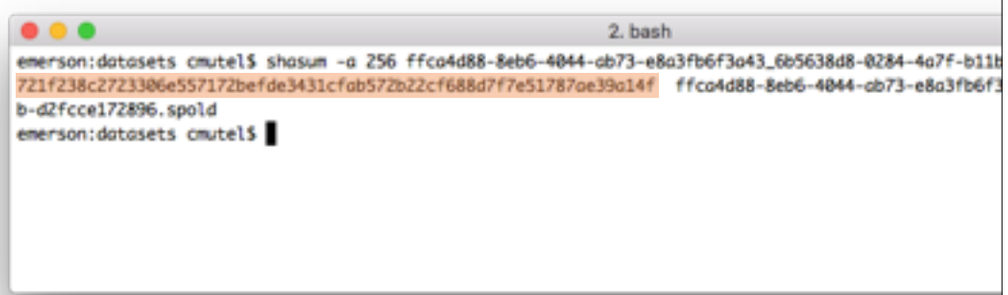
Vision

- JSON-LD format (OLCA schema)
- Lots of LCA software
  - Most do data collection
  - Most run in web browser
  - Calculations run in the cloud
  - Increased pressure for compatibility
- Sharing data is instant
  - Exact links to public and private datasets
- Dataset history is readable and passable
  - Separate bug fixes/updates/breaking changes
  - Authors produce official upgrade paths
- **Easier to share data than hide it**

Existing technologies

The web of names, hashes, and UUIDs

Identify activity datasets by their hashes



The web of names, hashes and UUIDs

Existing technologies

- Filesystem on the web
  - Distributed
- Locate files by hash
- Git-like diffing (still in progress)
- Need some metadata specifications

ipfs.io, dat data

Please email if interested:
cmutel@gmail.com

See also:
ocelot.space
brightwaylca.org
inventories.io