Applying metadata descriptors for finding data and for building LCA models

Fitness for purpose based autolinking in heterogenous data pools
Applying metadata descriptors for finding data and for building LCA models

Fitness for purpose based autolinking in heterogenous data pools

1. UN GLAD descriptors for assessing fitness for purpose
2. Application 1: Searching for best fitting datasets
3. Application 2: Linking in LCA product systems
   1. Concept
   2. Application case
   3. Next steps
4. Discussion: what would this concept change for LCA?
1 UN GLAD Descriptors for assessment of data quality
UN GLAD descriptors for assessing fitness for purpose (= data quality)

- Result of an international Working Group 2015 – 2017
- About 30 indicators that describe aspects of data quality for LCA inventory data sets
- Idea: capture content and LCA modelling details of LCA datasets
- Data quality is called fitness for purpose
- Recently implemented in an API and search engine, open source
  https://github.com/GreenDelta/glad-rest-service
- API integrated in a central website, globallcadataaccess.org
UN GLAD descriptors for assessing fitness for purpose (= data quality)

- Common structure for each descriptor:
  - **goal** (what the dataset ideally represents)
  - **representation / value** (what the dataset actually represents)
  - **conformance** (delta between goal and representation)
  - representation and conformance can be reviewed, the goal cannot be reviewed (since it is given by the application / user)
UN GLAD descriptors for assessing fitness for purpose (= data quality)

Ciroth, A., Arbuckle, P., Cherubini, E., Ugaya, C., Edelen, A.: Core meta-data descriptors and guidance on populating descriptors WG3 of the Global Life Cycle Data Access Network (GLAD), Release version 1.0 (internal version 11.3), June 2017
UN GLAD descriptors for assessing fitness for purpose (= data quality)

<table>
<thead>
<tr>
<th>API field</th>
<th>field format</th>
<th>field content</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>refId</td>
<td>String</td>
<td>The unique identifier of the data set *required</td>
<td>default</td>
</tr>
<tr>
<td>name</td>
<td>String</td>
<td>The name of the data set *required</td>
<td></td>
</tr>
<tr>
<td>dataSetUrl</td>
<td>String</td>
<td>A url to download the complete data set *required</td>
<td></td>
</tr>
<tr>
<td>category</td>
<td>String</td>
<td>The category of the data set. The value will be automatically build from the elements in the 'categories' field, concatenated with a slash (/). e.g. categories = ['Emission to air', 'Unspecified'] =&gt; category = 'Emission to air/Unspecified'</td>
<td></td>
</tr>
<tr>
<td>description</td>
<td>String</td>
<td>The description of the data set</td>
<td></td>
</tr>
<tr>
<td>technology</td>
<td>String</td>
<td>A description of the technology used in the data set</td>
<td></td>
</tr>
<tr>
<td>format</td>
<td>String</td>
<td>Enum: ECOSPOD1, ECOSPOD2, ILC, JSON-LD, OTHER, UNKNOWN</td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>String</td>
<td>The location of the data set</td>
<td></td>
</tr>
<tr>
<td>dataprovider</td>
<td>String</td>
<td>The name of the provider of the data set</td>
<td></td>
</tr>
<tr>
<td>supportedNomenclatures</td>
<td>String</td>
<td>The nomenclatures, the data set is compliant to</td>
<td></td>
</tr>
<tr>
<td>lciaMethods</td>
<td>[string]</td>
<td>A list of supported LCIA methods</td>
<td></td>
</tr>
<tr>
<td>categoryPaths</td>
<td>[string]</td>
<td>(Internally) used to build a tree like category structure. This value will be automatically calculated from the elements in the 'categories' field. e.g. categories = ['Emission to air', 'Unspecified'] =&gt; categoryPaths = ['Emission to air', 'Emission to air/Unspecified']</td>
<td></td>
</tr>
<tr>
<td>unspscPaths</td>
<td>[string]</td>
<td>(Internally) used to build a tree like structure for the unspsc code. This value will be automatically calculated from the 'unspscCode' field. e.g. unspscCode = '50454302' =&gt; unspscPaths = ['50', '5045', '504543', '50454302']</td>
<td></td>
</tr>
<tr>
<td>co2pePaths</td>
<td>[string]</td>
<td>(Internally) used to build a tree like structure for the co2pe code. This value will be automatically calculated from the 'co2peCode' field. e.g. co2peCode = '1.1.1' =&gt; co2pePaths = ['1', '1.1', '1.1.1']</td>
<td></td>
</tr>
<tr>
<td>processType</td>
<td>string</td>
<td>Enum: UNIT, PARTIALLY_AGGREGATED, FULLY_AGGREGATED, BRIDGE, UNKNOWN</td>
<td>Default: UNKNOWN</td>
</tr>
<tr>
<td>representativenessType</td>
<td>string</td>
<td>Enum: SCIENTIFIC, EXPERT_BASED</td>
<td>Default: EXPERT_BASED</td>
</tr>
<tr>
<td>modelingType</td>
<td>string</td>
<td>Enum: ATTRIBUTIONAL, CONSEQUENTIAL, UNKNOWN</td>
<td>Default: UNKNOWN</td>
</tr>
<tr>
<td>multifunctionalModeling</td>
<td>string</td>
<td>Enum: PHYSICAL, ECONOMIC, CAUSAL, SYSTEM_EXPANSION, NONE, UNKNOWN, NOT_APPPLICABLE</td>
<td>Default: NOT_APPPLICABLE</td>
</tr>
<tr>
<td>biogenicCarbonModeling</td>
<td>string</td>
<td>Enum: OMITTED, DISTINGUISHED, AGGREGATED, UNKNOWN, NOT_APPPLICABLE</td>
<td>Default: NOT_APPPLICABLE</td>
</tr>
<tr>
<td>endOfLifeModeling</td>
<td>string</td>
<td>Enum: CUT_OFF, PHYSICAL_APOS, ECONOMIC_APOS, SUBSTITUTION, OTHER, UNKNOWN, NOT_APPPLICABLE</td>
<td>Default: NOT_APPPLICABLE</td>
</tr>
<tr>
<td>waterModeling</td>
<td>string</td>
<td>Enum: AMOUNTS, AMOUNTS_AND_AVAILABILITY, AMOUNTS_AND_QUALITY, UNKNOWN, NOT_APPPLICABLE</td>
<td>Default: NOT_APPPLICABLE</td>
</tr>
<tr>
<td>infrastructureModeling</td>
<td>string</td>
<td>Enum: INCLUDED_AND_DISTINGUISHED, INCLUDED_AND_NOT_VISIBLE, NOT_INCLUDED, UNKNOWN, NOT_APPPLICABLE</td>
<td>Default: NOT_APPPLICABLE</td>
</tr>
</tbody>
</table>
## UN GLAD descriptors for assessing fitness for purpose (= data quality)

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>emissionModeling</td>
<td>string</td>
<td>A UNSPSC process code categorizing the data set</td>
</tr>
<tr>
<td>carbonStorageModeling</td>
<td>string</td>
<td>A CO2PE product code identifying the product of the data set</td>
</tr>
<tr>
<td>sourceReliability</td>
<td>string</td>
<td>The owner of the copyright of the data set if applicable</td>
</tr>
<tr>
<td>aggregationType</td>
<td>string</td>
<td>The license the data set is released under</td>
</tr>
<tr>
<td>reviewType</td>
<td>string</td>
<td>A contact person for information on the data set</td>
</tr>
<tr>
<td>reviewSystem</td>
<td>string</td>
<td>The categories of the data set as array (one entry per child category)</td>
</tr>
<tr>
<td>unspscCode</td>
<td>string</td>
<td>A list of the names of the reviewers of the data set</td>
</tr>
<tr>
<td>co2peCode</td>
<td>string</td>
<td>The start of the validity of the data set in milliseconds since 01/01/1970 (unix-time times 1000)</td>
</tr>
<tr>
<td>copyrightHolder</td>
<td>string</td>
<td>The end of the validity of the data set in milliseconds since 01/01/1970 (unix-time times 1000)</td>
</tr>
<tr>
<td>license</td>
<td>string</td>
<td>The year of the start of the validity of the data set, will be taken from validFrom if not set</td>
</tr>
<tr>
<td>contact</td>
<td>string</td>
<td>The year of the end of the validity of the data set, will be taken from validUntil if not set</td>
</tr>
<tr>
<td>categories</td>
<td>[string]</td>
<td>The latitude of the geography of the data set</td>
</tr>
<tr>
<td>reviewers</td>
<td>[string]</td>
<td>The longitude of the geography of the data set</td>
</tr>
<tr>
<td>validFrom</td>
<td>integer</td>
<td>The percentage of flows according to nomenclature</td>
</tr>
<tr>
<td>validFromYear</td>
<td>integer</td>
<td>The deviation in mass and energy balance</td>
</tr>
<tr>
<td>validUntil</td>
<td>integer</td>
<td>The percentage of variation coefficient, s/(arithm mean)</td>
</tr>
<tr>
<td>validUntilYear</td>
<td>integer</td>
<td>Indicates if the data set is copyright protected</td>
</tr>
<tr>
<td>latitude</td>
<td>number</td>
<td>Indicates if the data set is available for free</td>
</tr>
<tr>
<td>longitude</td>
<td>number</td>
<td>Indicates if the data set can be downloaded from the given dataSetUrl without further login</td>
</tr>
<tr>
<td>completeness</td>
<td>number</td>
<td>Default: NOT_APPLICABLE</td>
</tr>
<tr>
<td>amountDeviation</td>
<td>number</td>
<td>Default: ESTIMATED_UNQUALIFIED</td>
</tr>
<tr>
<td>representativenessValue</td>
<td>number</td>
<td>Default: NONE</td>
</tr>
<tr>
<td>copyrightProtected</td>
<td>boolean</td>
<td>Default: NOT_APPLICABLE</td>
</tr>
<tr>
<td>free</td>
<td>boolean</td>
<td>Default: NOT_APPLICABLE</td>
</tr>
</tbody>
</table>
UN GLAD descriptors for assessing fitness for purpose (= data quality)

```
infrastructureModeling: string

Enum: INCLUDED_AND DISTINGUISHED, INCLUDED_AND_NOT_VISIBLE, NOT_INCLUDED, UNKNOWN, NOT_APPLICABLE
```

descriptor name

descriptor representation: format and content
2 Application case 1: finding best fitting datasets
Finding datasets best fitting for purpose (i.e., of best quality), application case 1

- Searching sugar-related datasets in LCA databases

- Sugar cane, sugar beet (Albania)
  - Databases: Social Hotspots Database
  - Category: Social Hotspots DB/Albania
  - Version (internal): 00.00.000  Location: Albania

- Sugar cane, sugar beet (Armenia)
  - Databases: Social Hotspots Database
  - Category: Social Hotspots DB/Armenia
  - Version (internal): 00.00.000  Location: Armenia

- Sugar cane, sugar beet (Angola)
  - Databases: Social Hotspots Database
  - Category: Social Hotspots DB/Angola
  - Version (internal): 00.00.000  Location: Angola
Finding datasets best fitting for purpose (i.e., of best quality), application case 1

- User specifies what he or she needs (→ “goal”)
- Search engine has access to datasets and to descriptors of datasets (to the “value and representation” part)
- Search engine applies conformance assessment for each descriptor
- All conformance results are merged to one fitness for purpose index
- Datasets are presented to users as search result list
- User can then select and download the dataset in a desired exchange format
3 Application case 2: linking in LCA product systems
LCA Product systems

- LCA Product systems are linked life cycles

Product systems: “collection of unit processes with elementary and product flows, performing one or more defined functions, and which models the life cycle of a product” [ISO 14040, 3.28]
LCA Product systems
LCA Product systems

One example connection: sugar beet production in Germany needs phosphate fertiliser.
In order to establish this connection, a process needs to be found that ideally:

- fits to goal and scope of the study (and product system)
  - same reference time
  - same allocation model
  - same way to model biogenic carbon
  - ...
- delivers the specific product needed by the sugarbeet process
LCA Product systems

In order to establish this connection, a process needs to be found that ideally:

• fits to goal and scope of the study (and product system)

• delivers the specific product needed by the sugarbeet process

→ i.e., a process needs to be found that best fits for purpose (!)

→ we will call this approach *fitness for purpose autolinking*, FFPA
LCA Product systems

Linking “the next” process in a product system is a bit like taking this or the other road when driving: When you take one direction, this influences the next “available” streets.
LCA Product systems, sugarbeet example

sugarbeet production
input
1 manure
2 N-fertiliser
3 P-fertiliser
...

product system

1: manure
2: N-fertiliser biogenic c: distinguish
   reference time: 2015
   region: Germany
   ..

2: N-fertiliser
   reference (from sugarbeet prod.)
   biogenic c: distinguish
   reference time: 2015
   region: Germany
   ..
technology (process & product)
   diammonium phosphate
## LCA Product systems, sugarbeet example

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
<th>Biogenic</th>
<th>Reference Time</th>
<th>Region</th>
<th>Technology (Process &amp; Product)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>N-fertiliser (from sugarbeet prod.)</td>
<td>distinguish</td>
<td>2015</td>
<td>Germany</td>
<td>Diammonium phosphate</td>
</tr>
<tr>
<td>Option A</td>
<td>Calcium ammonium nitrate (CAN, solution), production mix, at plant, technology mix, nitrogen content 26.5% (United States)</td>
<td>distinguish</td>
<td>2013</td>
<td>US</td>
<td>Calcium ammonium nitrate, production mix</td>
</tr>
<tr>
<td>Option B</td>
<td>Urea ammonium nitrate (UAN), production mix, at plant, technology mix, nitrogen content 30% (Germany)</td>
<td>distinguish</td>
<td>2013</td>
<td>Germany</td>
<td>Urea ammonium nitrate, production mix</td>
</tr>
<tr>
<td>Option C</td>
<td>Nitrogen fertiliser, use of (Switzerland)</td>
<td>aggregated</td>
<td>2011</td>
<td>Switzerland</td>
<td>Nitrogen fertiliser, unspecified [, production mix]</td>
</tr>
</tbody>
</table>
Linking in LCA product systems, next steps

• Provide descriptors for important LCA data sources (also for GLAD, for the search engine)

• Implement the search in LCA software (e.g., openLCA):
  • linking algorithm
  • results display, “data quality” of the links in a product system (instead of the quality of the process)

• Refine concept (i.e., there is a propagation of fitness for purpose in linking: a bad fit in step i cannot entirely be compensated by a good fit in step i+1)
3 Discussion: what would this autolinking change
Discussion points: what would FFPA change?

- status now: existing databases, some (ecoinvent e.g.) with foreseen linking, struggling to capture more data
- combination of different databases hard to decide, typically specific user decisions
- FFPA:
  - agnostic to specific databases, can use any data source available
  - allows to “throw in” new data sources in a data pool, data is considered once it improves the linking, and in the end quality of the model
  - promise: more complete, more realistic model, fitting (much!) better to what the user needs
Discussion points: what would FFPA change?

- FFPA, continued:
  - best suited for unit processes but works with system processes too (fewer linking decisions)
  - works also with already well-linked databases
  - a product system linking is then like a transportation network:
    - motorways (well-linked databases)
    - railways (system processes, a route change is not easily possible)
    - and smaller streets (everything else)
Discussion points: what would FFPA change?

Uber’s Kepler application: commute patterns in the UK, http://kepler.gl

GreenDelta  A. Ciroth, S. Greve: GLAD autolinking, Sept 25, 2018
FFPA status & call for participation

- Descriptors now in 1st version, more are proposed by the metadata WG; they need to be tested.
- E.g. in the Ocelot project, other aspects have been considered but only for ecoinvent (production volume e.g.)
- (1) method, (2) databases, and (3) tool support are needed to make FFPA really powerful.
- We plan to extend the idea and extend openLCA and the Nexus databases and our LCA data machine to include the descriptors and to support the FFPA autolinking.
- Any collaboration is welcome!
Thank you!

Contact: Dr. Andreas Ciroth
GreenDelta GmbH
Müllerstrasse 135, 13349 Berlin, Germany
ciroth@greendelta.com
www.greendelta.com