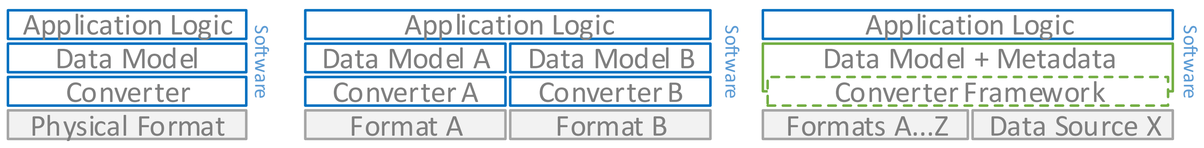
University of Stuttgart
Germany

ICARUS2 SDK

A Lightweight Modeling Middleware for Corpus Processing

Markus Gärtner, Jonas Kuhn

Middleware Approach



- shifts unification from entire format-stacks (left and center graphics) to a dedicated middleware framework (green part of right graphic)
- aimed at exploration and visualization tools that require unified access to very rich and diverse corpus resources without losing linguistic specifics
- modeling framework for in-memory representation of corpus resources
- actual modeling task split into a graph-like data model and a metadata framework for describing corpus composition and linking to linguistic categories

Metadata Model

- used to describe composition and dependencies of a corpus:
  - layers for expressing structure, relations and annotations
  - types, tagsets or other constraints for annotation values
  - grouping mechanisms (`context`, `layer group`) for layers to account for e.g. multiple physical sources composing a single corpus resource
  - dependencies between layers, layer groups and contexts
- programing language independent, XML as default serialization format
- supports linking of elements such as layers to linguistic categories
- template support for better reusability

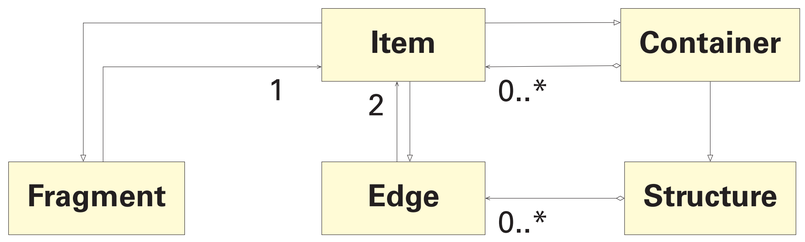
```
<annotation id="common.tags.stts" name="
STTS-Tagset", valueType="string">
 <values>
  <value name="Adjective">JJ</value>
  <value name="Noun">NN</value>
  <value name="Determiner">DT</value>
  <value name="Verb, gerund">VBG</value>
  <value name="Verb, 3rd sg">VBZ</value>
  [...]
 </values>
</annotation>
```

```
<corpus id="my.simple.corpus">
 <context foundation="token">
  <layerGroup primary="token">
   <itemLayer id="token"/>
   <itemLayer id="sentence">
    <baseLayer layerId="token"/>
    <container containerType="span"/>
   </itemLayer>
   <annotationLayer id="content">
    <baseLayer layerId="token"/>
    <annotation key="id" valueType="int"/>
    <annotation key="form"/>
    <annotation key="lemma"/>
    <annotation key="pos" templateId="common.tags.stts">
     <!-- Inherited from template -->
    </annotation>
   </annotationLayer>
  </layerGroup>
 </context>
</corpus>
```

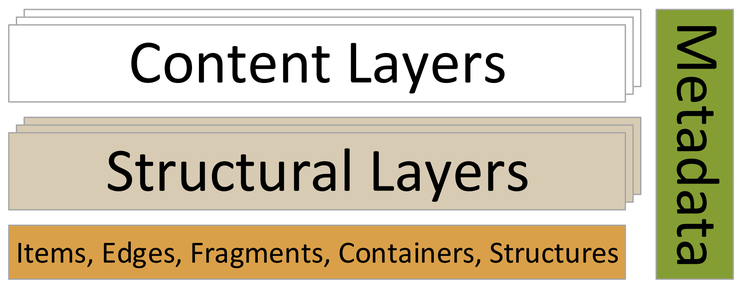
Example metadata describing a reusable template for a part-of-speech tagset (top) and a shallow corpus (right). The latter also makes use of the previously defined tagset template.

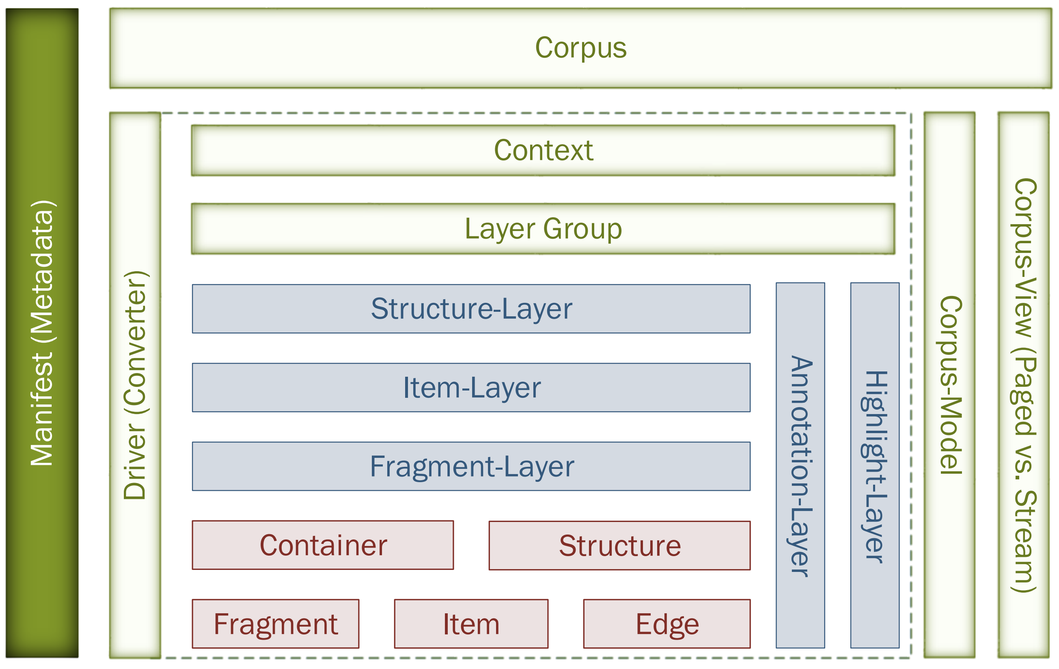
Data Model

- inspired by classic general purpose graph models
- designed to model arbitrary corpus compositions
- basic set of generic building blocks:



- separation of corpus structure (segmentation, hierarchies and relational structures) and the associated content (annotations)
- all elements linked to descriptive metadata for building more informed systems on top of them:





Hierarchy and interaction of different framework members: metadata (green), atomic building blocks (red), organizational layers (blue) and surrounding management structures (white).

Design Features

- multiple access modes to a corpus object (paged vs. streamed)
- notification system to propagate information about changes
- editable vs. static corpora with edit history support
- individual units in a corpus addressable via unique ids or the position in their respective host layers
- extensible set of natively supported annotation types
- scalable both horizontally (number of elements in a corpus) and vertically with the number of annotation layers
- separation of data model and descriptive metadata provides media independence and increased flexibility

Availability

- modeling framework implemented in Java
- open-source software available on GitHub
  `https://github.com/ICARUS-tooling`
- metadata and additional documentation available in the context of CLARIN:
  `http://hdl.handle.net/11022/1007-0000-0007-C636-D`

www.ims.uni-stuttgart.de

This work was funded by the German Federal Ministry of Education and Research (BMBF) via CLARIN-D, No. 01UG1120F and the German Research Foundation (DFG) via the SFB 732, project INF.