**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
  - Programmable language independent, XML as default serialization format
  - Supports linking of elements such as layers to linguistic categories
  - Template support for better reusability

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.

**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

---

**ICARUS 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)
- Aims 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

**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)

---

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.