


DKPro Core

Software components for NLP


M.Sc. Pedro Santos – Ubiquitous Knowledge Processing
Lab

[DKPRO](#) [CORE](#) [DOWNLOADS](#) [DOCUMENTATION](#) [ISSUES](#) [SOURCE](#) [CONTACT](#) [ABOUT](#)


 **DKPro Core**

A collection of software components for natural language processing (NLP) based on the Apache UIMA framework.


Many NLP tools are already freely available in the NLP research community. DKPro Core provides Apache UIMA components wrapping these tools (and some original tools) so they can be used interchangeably in UIMA processing pipelines. DKPro Core builds heavily on `uimaFIT` which allows for rapid and easy development of NLP processing pipelines, for wrapping existing tools and for creating original UIMA components. [More](#)




Components
Find out more about our bundled components.



Models/Languages
Various models covering different languages accompany the components.




Formats
Reading and writing various formats is just one line of code away.




Typesystem
Our typesystem is comprehensive, yet simple.


Latest release: 1.7.0 (2014-11-28)



DKPro with Java
The original flavour. Use DKPro in your Java projects.



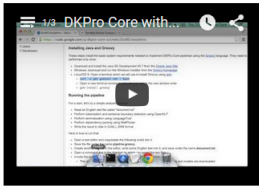
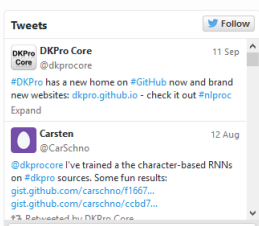
DKPro with Groovy
Create self-contained scripts using DKPro and Groovy!



DKPro with Jython
Easily integrate DKPro into your python projects!

How to cite

Many of the wrapped third-party components and the models used by them should be cited individually. We currently do not provide a comprehensive overview over citable publications. We encourage you to track down citable publications for these dependencies. However, you might find pointers to some relevant publications in the Model overview of the DKPro Core release you are using or in the JavaDoc of individual components.

Agenda

- What is a pipeline?
- Working with annotations
 - What is a type system?
 - What is the Common Analysis Structure (CAS)?
- Working with components
 - What is a reader?
 - What is an analysis engine?
 - What is a writer? (aka consumer)
- DKPro Core component collection

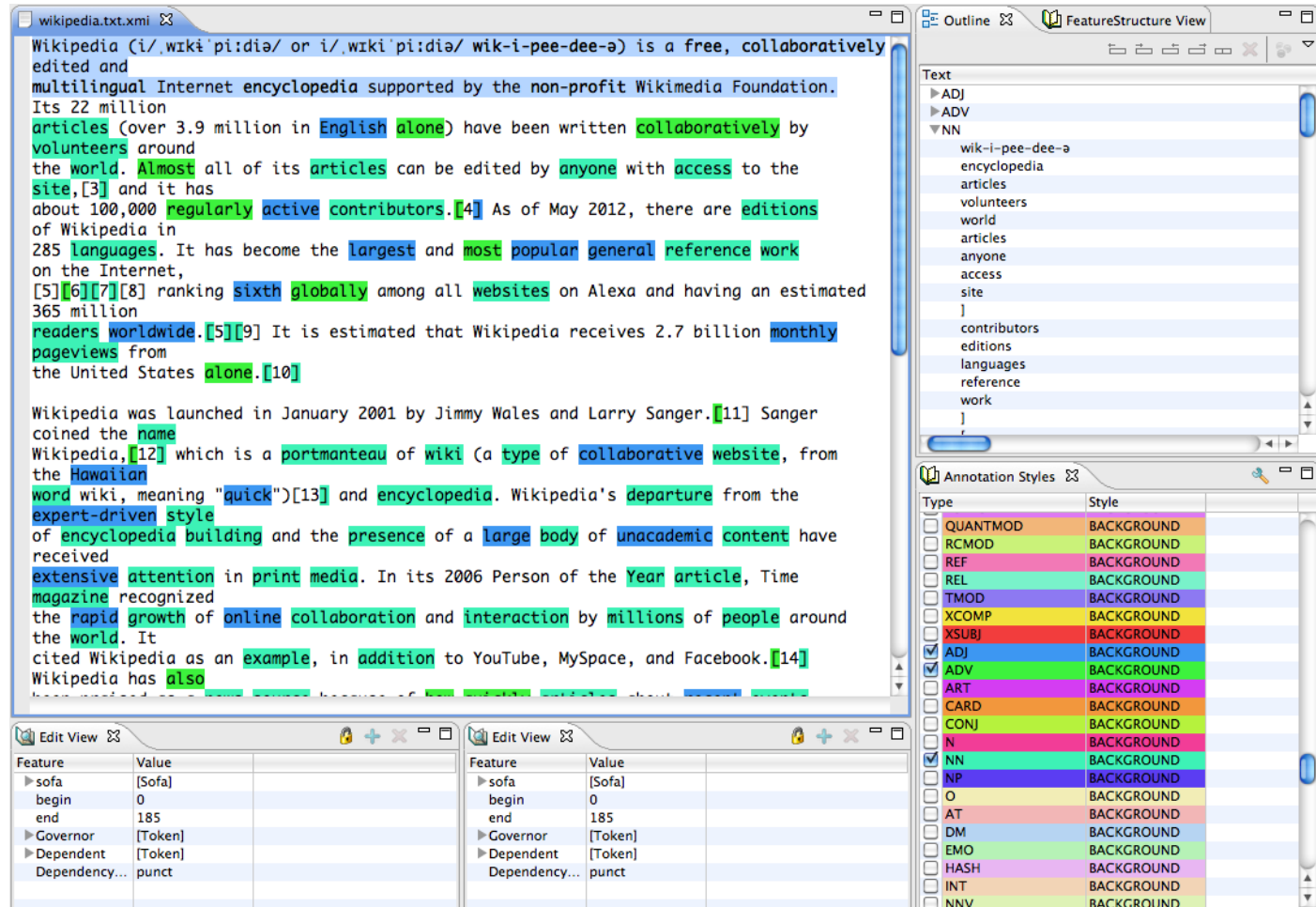


What is UIMA?

- Component-based architecture
 - Analysis of **unstructured data**
 - Structure from the unstructured data
- How?
 - Like an assembly line...
 - Raw material
 - Refinement, step by step
 - Nice car in the end



Output Example (UIMA Annotation Editor)



The screenshot displays the UIMA Annotation Editor interface. The main window shows a text document titled "wikipedia.txt.xml" with the following text:

Wikipedia (i/ˌwɪkɪˈpiːdiə/ or i/ˌwɪkiˈpiːdiə/ wik-i-pee-dee-ə) is a free, collaboratively edited and multilingual Internet encyclopedia supported by the non-profit Wikimedia Foundation. Its 22 million articles (over 3.9 million in English alone) have been written collaboratively by volunteers around the world. Almost all of its articles can be edited by anyone with access to the site,[3] and it has about 100,000 regularly active contributors.[4] As of May 2012, there are editions of Wikipedia in 285 languages. It has become the largest and most popular general reference work on the Internet,[5][6][7][8] ranking sixth globally among all websites on Alexa and having an estimated 365 million readers worldwide.[5][9] It is estimated that Wikipedia receives 2.7 billion monthly pageviews from the United States alone.[10]

Wikipedia was launched in January 2001 by Jimmy Wales and Larry Sanger.[11] Sanger coined the name Wikipedia,[12] which is a portmanteau of wiki (a type of collaborative website, from the Hawaiian word wiki, meaning "quick") [13] and encyclopedia. Wikipedia's departure from the expert-driven style of encyclopedia building and the presence of a large body of unacademic content have received extensive attention in print media. In its 2006 Person of the Year article, Time magazine recognized the rapid growth of online collaboration and interaction by millions of people around the world. It cited Wikipedia as an example, in addition to YouTube, MySpace, and Facebook.[14] Wikipedia has also been named one of the most influential websites because of its impact on the world.

The interface includes several panels:

- Outline:** A tree view showing the document structure, including "Text", "ADJ", "ADV", "NN", and "NP".
- FeatureStructure View:** A panel for viewing and editing feature structures.
- Annotation Styles:** A table for defining annotation styles.
- Edit View:** Two panels for editing annotations, showing features like "sofa", "begin", "end", "Governor", "Dependent", and "Dependency..." with their corresponding values.

Type	Style
<input type="checkbox"/> QUANTMOD	BACKGROUND
<input type="checkbox"/> RCMOD	BACKGROUND
<input type="checkbox"/> REF	BACKGROUND
<input type="checkbox"/> REL	BACKGROUND
<input type="checkbox"/> TMOD	BACKGROUND
<input type="checkbox"/> XCOMP	BACKGROUND
<input type="checkbox"/> XSUBJ	BACKGROUND
<input checked="" type="checkbox"/> ADJ	BACKGROUND
<input checked="" type="checkbox"/> ADV	BACKGROUND
<input type="checkbox"/> ART	BACKGROUND
<input type="checkbox"/> CARD	BACKGROUND
<input type="checkbox"/> CONJ	BACKGROUND
<input type="checkbox"/> N	BACKGROUND
<input checked="" type="checkbox"/> NN	BACKGROUND
<input type="checkbox"/> NP	BACKGROUND
<input type="checkbox"/> O	BACKGROUND
<input type="checkbox"/> AT	BACKGROUND
<input type="checkbox"/> DM	BACKGROUND
<input type="checkbox"/> EMO	BACKGROUND
<input type="checkbox"/> HASH	BACKGROUND
<input type="checkbox"/> INT	BACKGROUND
<input type="checkbox"/> NNV	BACKGROUND

Apache UIMA™ – Some history

- 2003 – David Ferrucci and Adam Lally paper
 - *Accelerating corporate research in the development, application and deployment of human language technologies*
- 2004 – IBM alphaWorks project
 - IBM LanguageWare
- 2006 – Apache Incubator project
- 2009 – OASIS Standard
- 2010 – Full Apache project
- 2010 – IBM's *Watson* Jeopardy Challenge



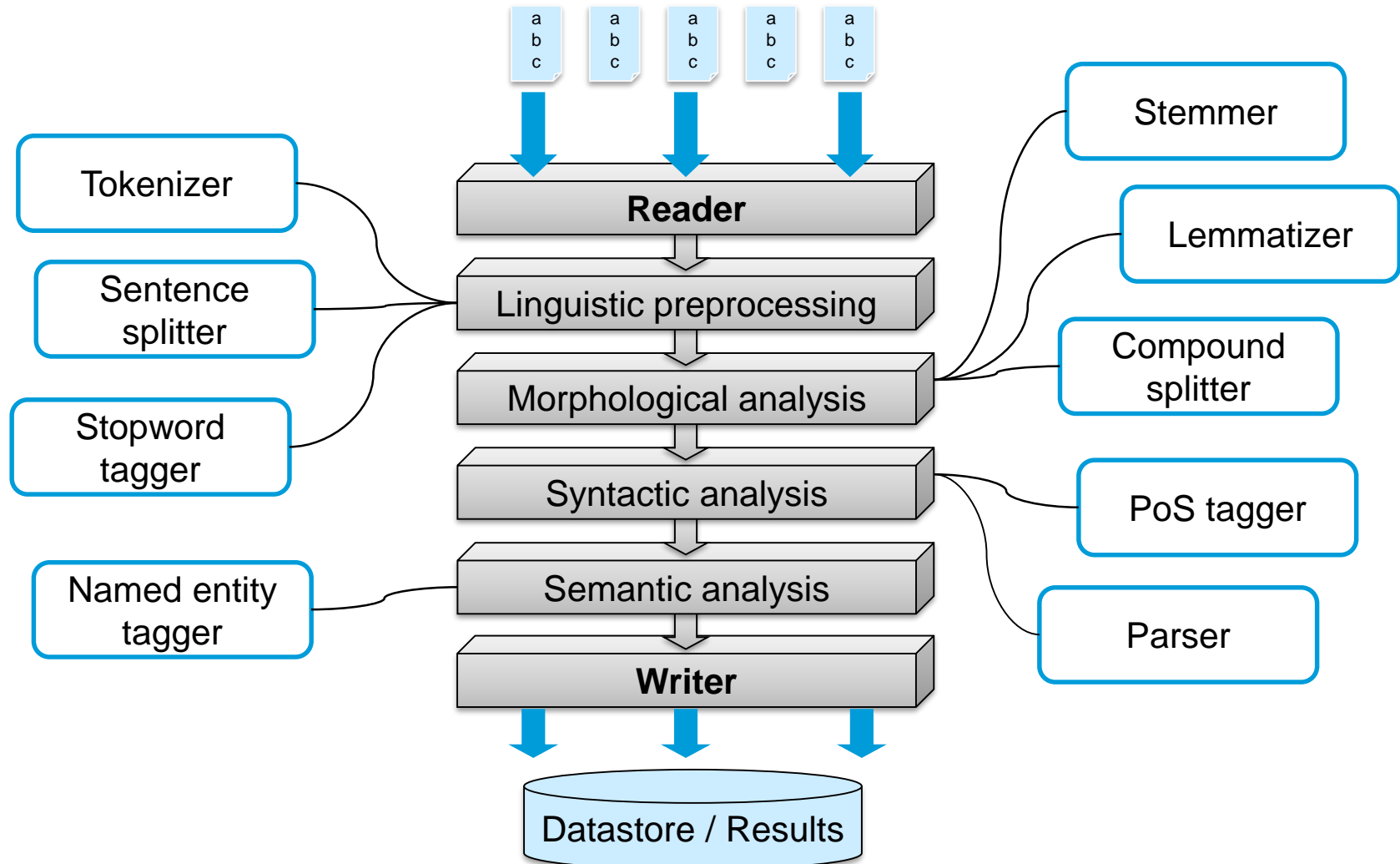
<http://uima.apache.org>

Pipelines



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Pipeline Architecture



Component – Collection Reader

- Empty data structure (CAS) → Reader
- Reader → Text (SofA) and Meta-Data (e.g. language)

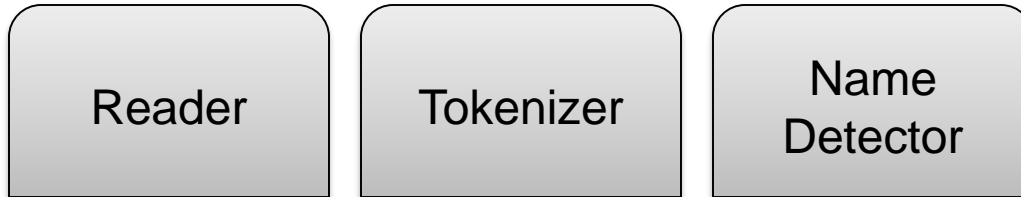
Reader

CAS

SofA	Language:	Latin
	DocumentText:	Ubi est Cornelia? Subito Marcus vocat: „Ibi Cornelia est, ibi stat!“

Component – Analysis Engine

- Structure → Analysis Engine (AE)
- Analysis Engine → Annotation

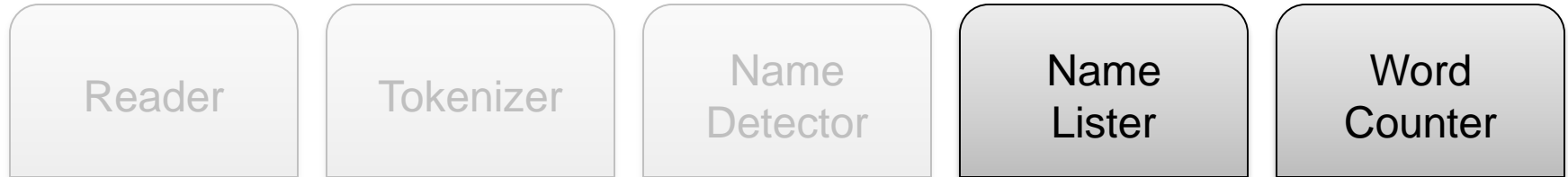


CAS

SofA Language: Latin
DocumentText: Ubi est Cornelia?
Subito Marcus vocat:
„Ibi Cornelia est, ibi stat!“

Token(0, 3) Token(4, 7) Token(8,16) ...
Name(8, 16) Name(25, 31) ...

- Annotation -> CAS Consumer



CAS

SofA Language: Latin
DocumentText: Ubi est Cornelia?
 Subito Marcus vocat:
 „Ibi Cornelia est, ibi stat!“

Token(0, 3) Token(4, 7) Token(8,16) ...
Name(8, 16) Name(25, 31) ...

Cornelia
Marcus

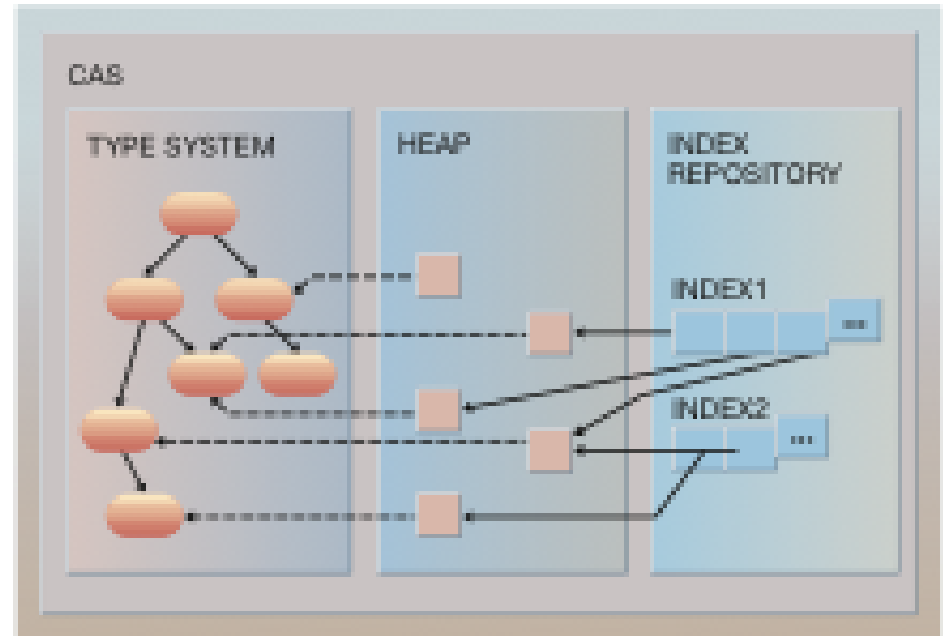
11 words
8 unique words



UIMA Data Structures

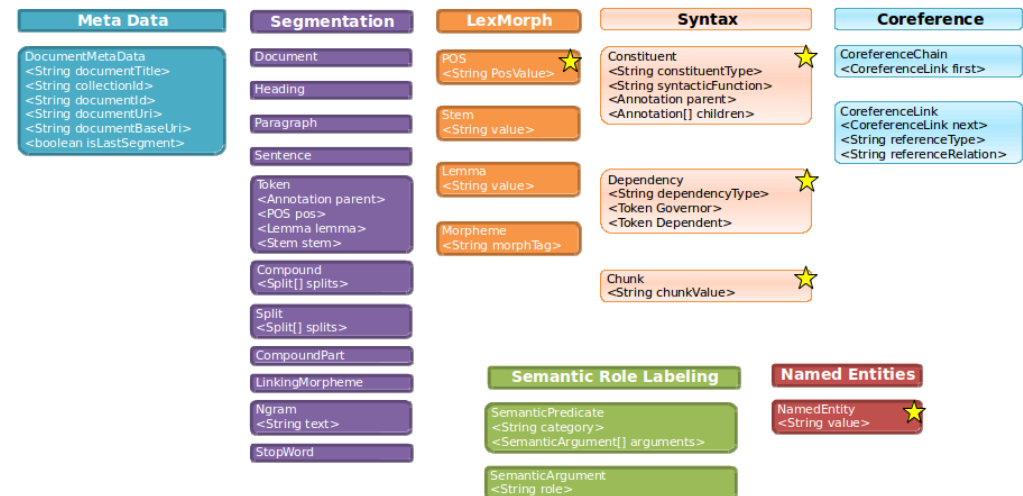
Common Analysis System (CAS)

- Access to primary data
- Secondary data storage
 - a.k.a. Annotations
- In-memory database
 - Annotation types = tables
 - Indexes



- Platform-independent specification
- Object-oriented type system:
 - Type → class
 - Feature → class member
 - Feature Structure → instance
 - Single inheritance
 - Sub-type polymorphism
 - No methods or encapsulation
- Primitive types: integer, float, boolean, string
- Built-in complex types: arrays, lists, Annotation
- Communication contract

DKPro Core Type System (Top Level)



★ For these types, DKPro Core provides several specialized subtypes, e.g. *NP* for noun phrase constituents or *Location* for places.

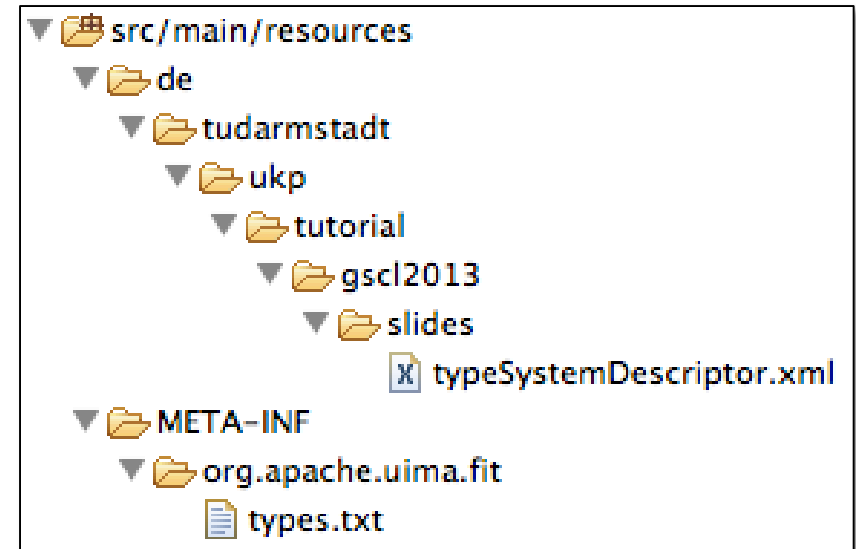
Java + CAS = JCas

- JCas = CAS types into the Java type system
- JCasGen
 - Java classes from XML type system descriptor
 - Token.java – feature structure wrapper with getters and setters
 - Token_type.java – type wrapper (cf. Java 'Class' class)
- JCas wrappers cannot be used stand-alone



uimaFIT type system detection

- No explicit loading/creation of type system
- Type system detection mechanism
- Types defined in XML descriptor files
- Scanning of classpath for type system descriptor files



Type System Editor (Eclipse)

- JCasGen → UIMA types available as Java Classes

Type System Definition

▼ **Types (or Classes)**

The following types (classes) are defined in this analysis engine descriptor.
The grayed out items are imported or merged from other descriptors, and cannot be edited here. (To edit them, edit their source files).

Type Name or Feature Name	SuperType or Range	Element Type
<input type="checkbox"/> de.tudarmstadt.ukp.tutorial.gsc12013.slides.Token	uima.tcas.Annotation	
length	uima.cas.Integer	
de.tudarmstadt.ukp.tutorial.gsc12013.slides.Name	uima.tcas.Annotation	
de.tudarmstadt.ukp.tutorial.gsc12013.slides.Sentence	uima.tcas.Annotation	
de.tudarmstadt.ukp.tutorial.gsc12013.slides.Paragraph	uima.tcas.Annotation	

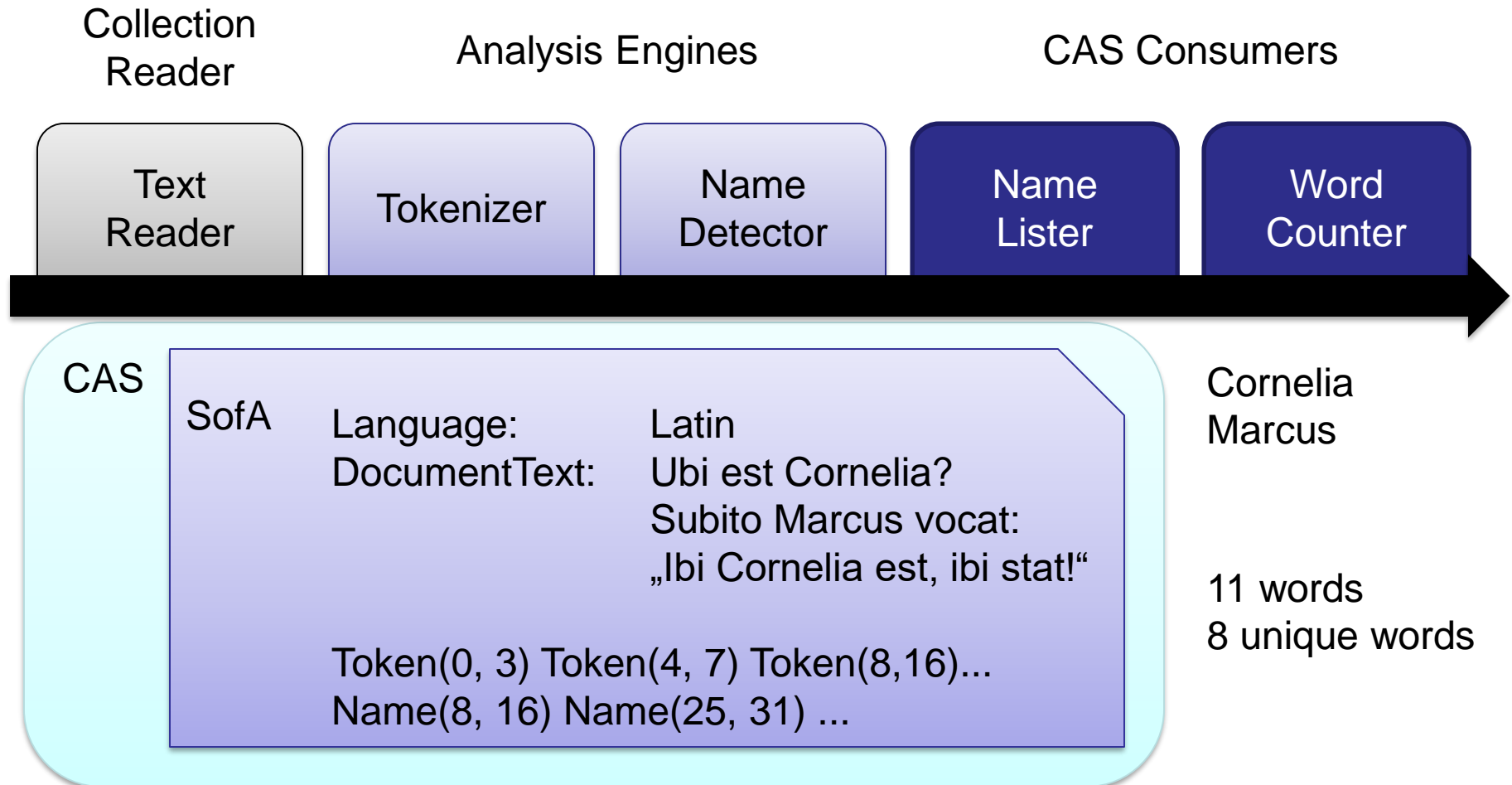
Add Type
Add...
Edit...
Remove
Export...
JCasGen
☐ limited

Components



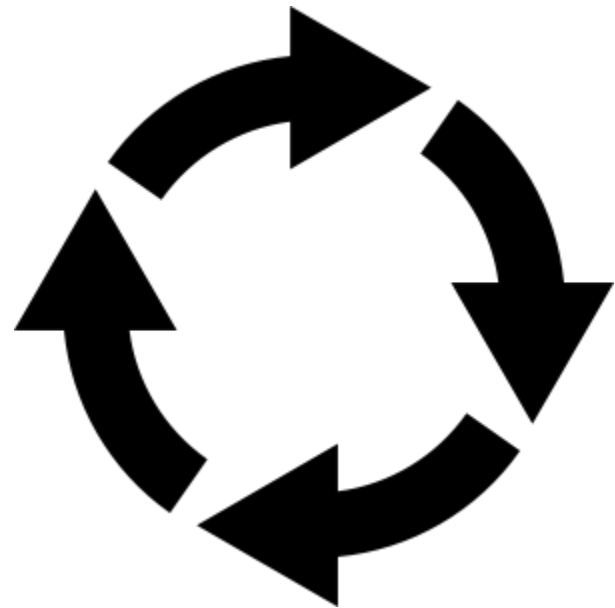
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Components



API – Life-Cycle Events

- Component life-cycle events
 - initialize()
 - reconfigure()
 - destroy()
- Processing life-cycle events
 - collectionProcessComplete()
 - batchProcessComplete()
- Other
 - typeSystemInit()



API – Processing Methods

- **CollectionReader**
 - hasNext()
 - getNext()
 - getProgress()
- **AnalysisEngine**
 - process()
- **CasConsumer**
 - process()

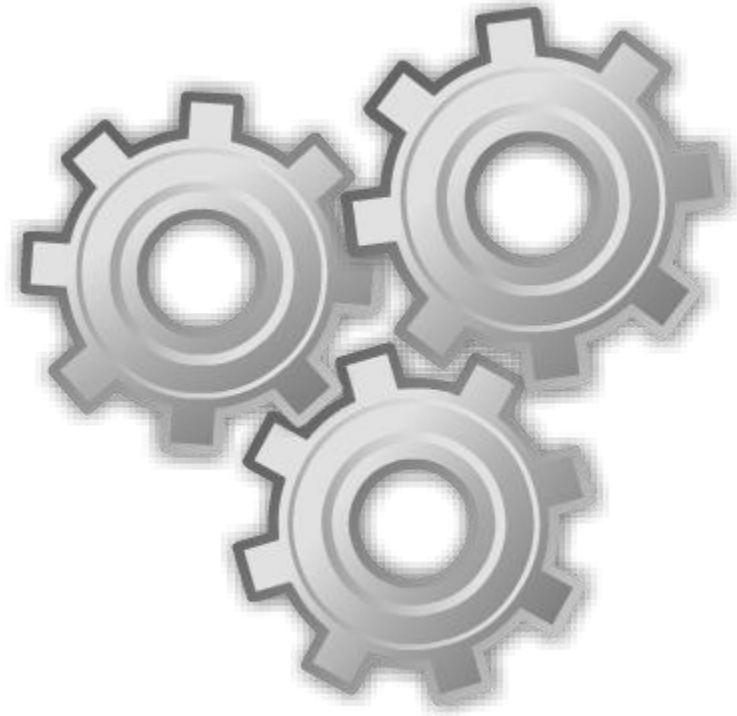


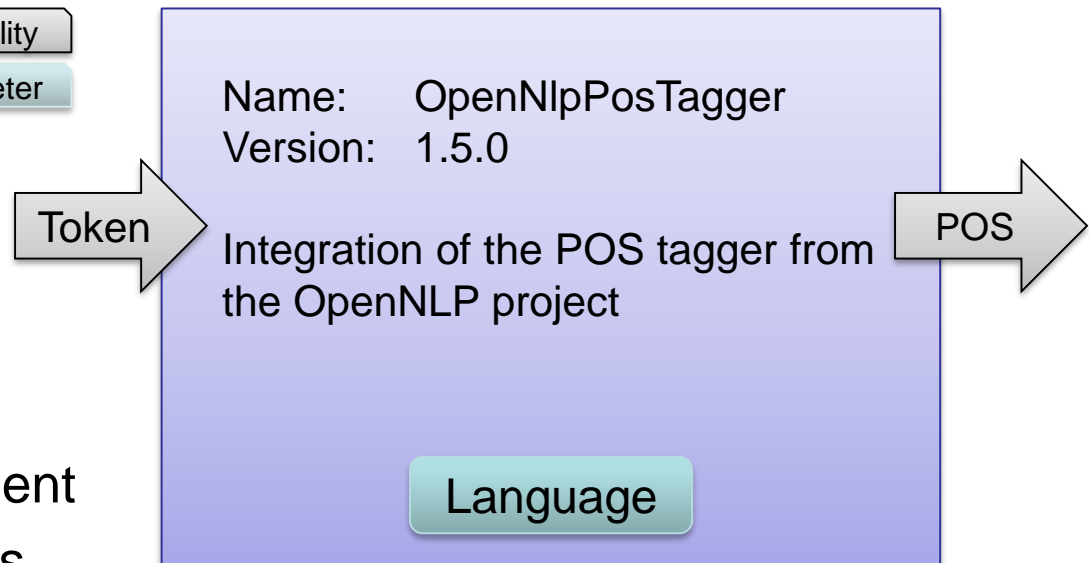
Figure: Analysis Engine Descriptor

- Name
- Version
- Vendor
- Type system
- Parameters
- Capabilities
- Indexes
- Resources
- Single- / multiple deployment
- Delegate Analysis Engines
- Flow control
- ... a few more

Legend

Capability

Parameter



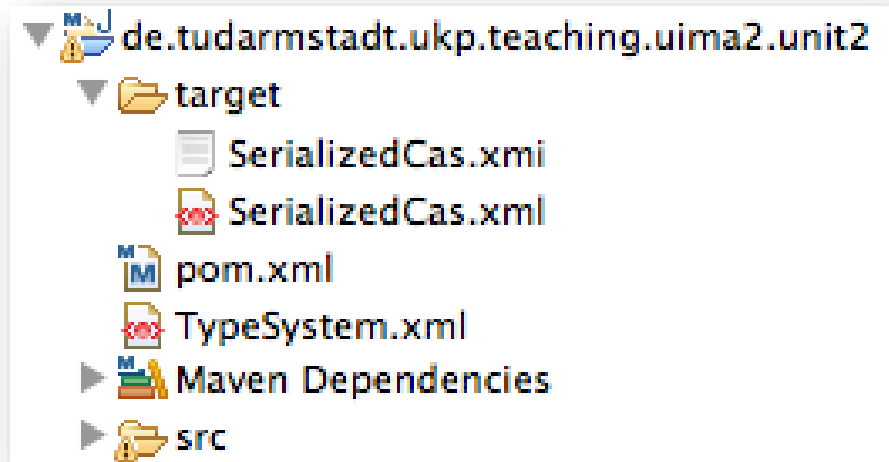
XML Descriptors – Pro & Contra

- Pro
 - “Officially preferred” form of configuration from UIMA components/resources
 - Widely supported by UIMA tooling
 - XML elements ↔ Java classes
- Contra
 - Mix declaration/documentation and configuration
 - Not included when refactoring code
 - No convenient API (remedy: uimaFIT factories)



Persisting and loading a CAS

- Available serialization formats
 - XCAS
 - XML format
 - XMI
- Type-system definition not included!
- Tip
 - Persist type system as “TypeSystem.xml” at project root
 - Open and XMI file in that project with the CAS Editor

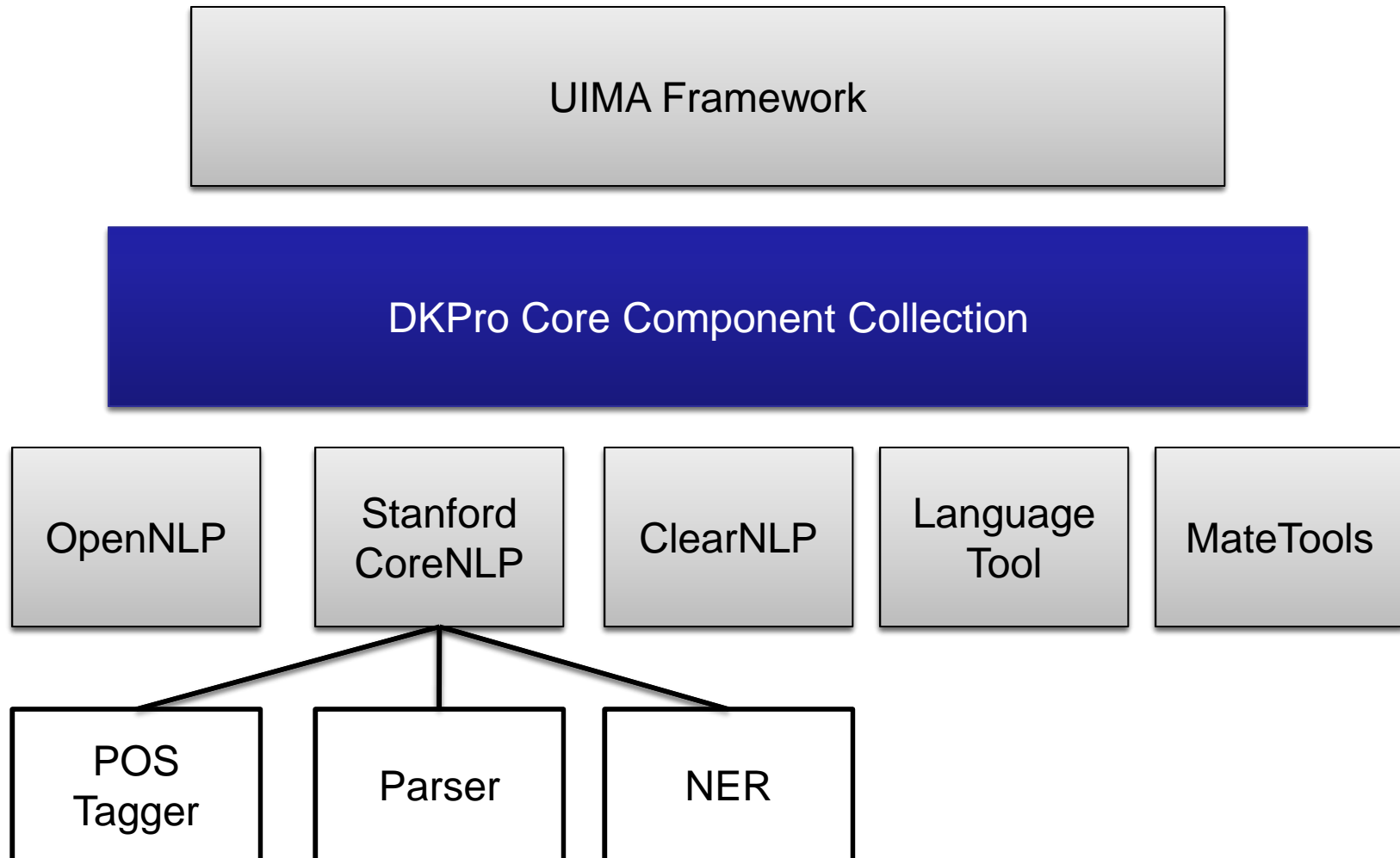


DKPro Core Component Collection

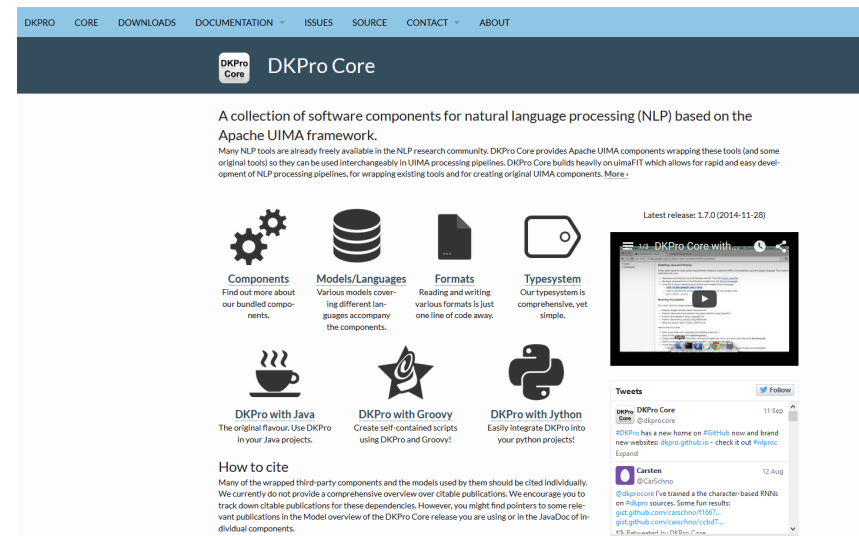


TECHNISCHE
UNIVERSITÄT
DARMSTADT

What's a component collection



- Integration framework
 - Processing: tools and models
 - Primary data: corpora
 - Auxiliary data: other language resources (e.g. lexical resources)
- Primarily integration of existing work, not original work
- Contribution: integration itself
- Open Source under Apache Software License & GNU Public License



<https://dkpro.github.io/dkpro-core/>

▪ **Simplicity**

- Common data types
- Common set of parameters
- Sensible parameters defaults for minimal need for configuration
- Convenient deployment of components and resources
- Compose powerful pipelines with a few lines of code

▪ **Modularity**

- Use only what you need

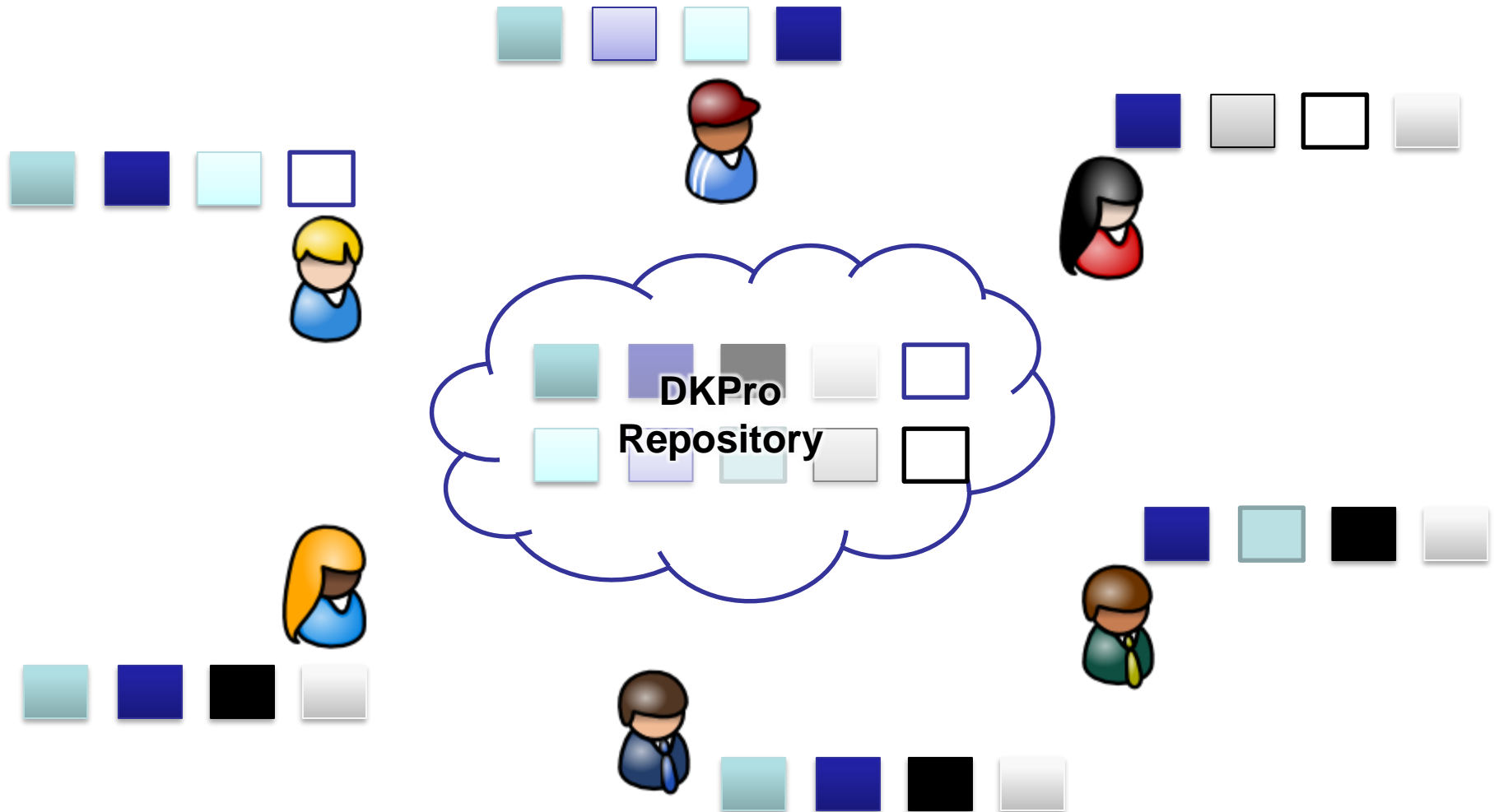
▪ ***Stuff has to “just work”, everywhere.***

- **Flexibility**

- Parameters override for fine-grained control
- Data types extension with custom fields
- Type mappings customization

▪ ***Stuff has to “just work”, everywhere.***

Managing Deployment



UKP OSS Component Repository

Publish component

Overview

Artifact

Group Id:

Artifact Id*:

Version:

Packaging:

Parent

Group Id*:

Artifact Id*:

Version*:

Relative Path:

Properties

Project

Name:

URL:

Description:

Inception:

Organization

SCM

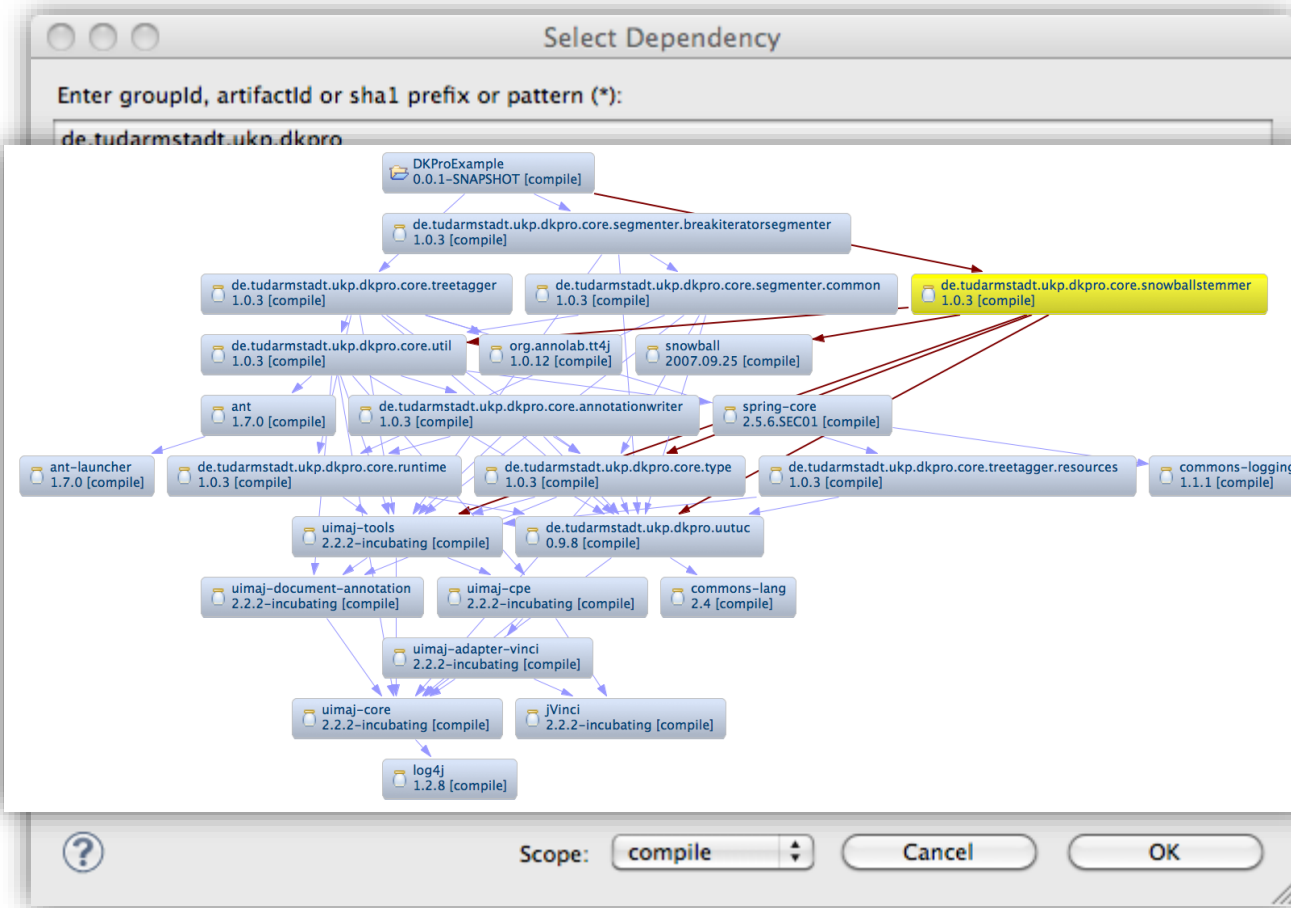
Issue Management

Continuous Integration



UKP OSS Component Repository Retrieving components

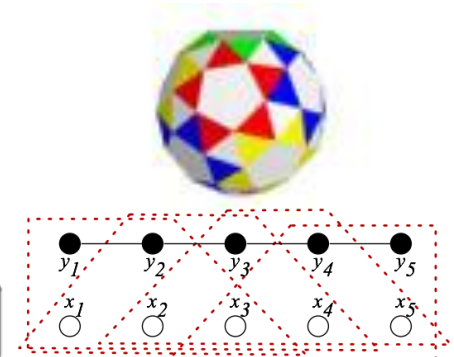
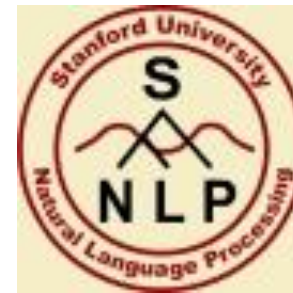
Component Repository



Tools and formats

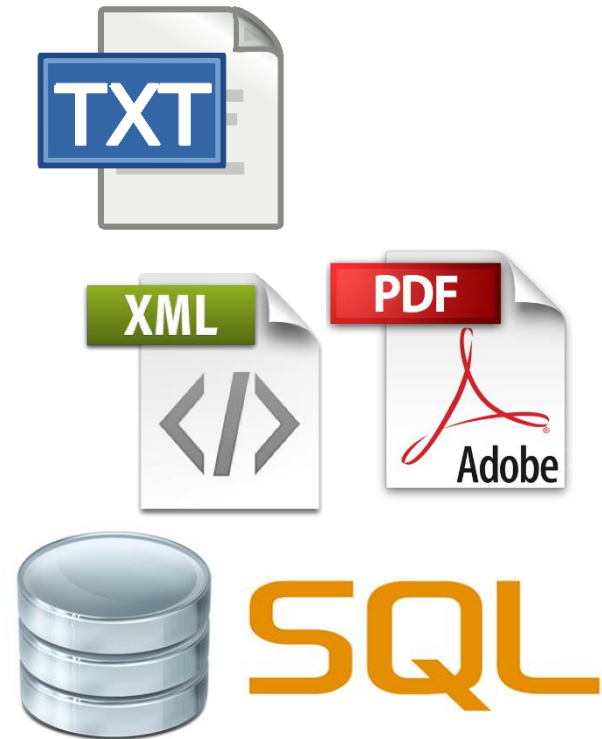
■ Integrated tools

- Stanford NLP
- OpenNLP
- Mate-Tools
- ClearNLP
- LanguageTool
- TreeTagger
- JWordSplitter
- Snowball Stemmer
- TextCat
- MaltParser
- MstParser
- BerkeleyParser
- ...



▪ Supported formats

- Text
- PDF
- TIGER XML
- TEI XML
- BNC XML
- Negra Export
- SQL Databases
- Google web1t n-grams
- ...

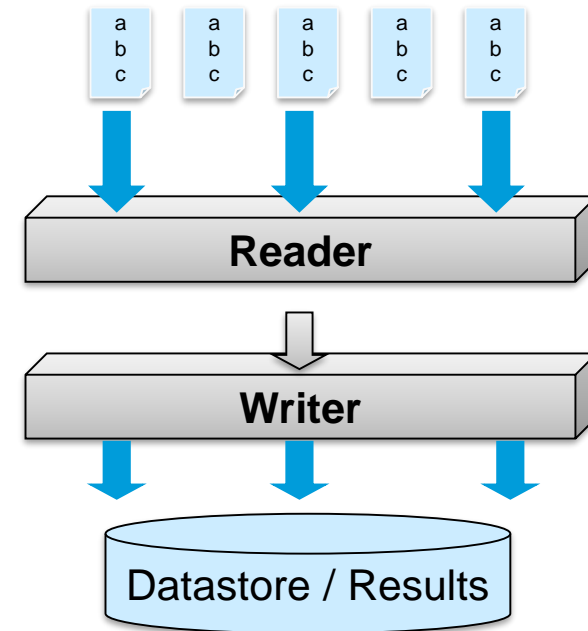


▪ Common parameters

- Source / target location
- Source / target encoding
- ANT-like patterns (for readers)
- Language (for readers)

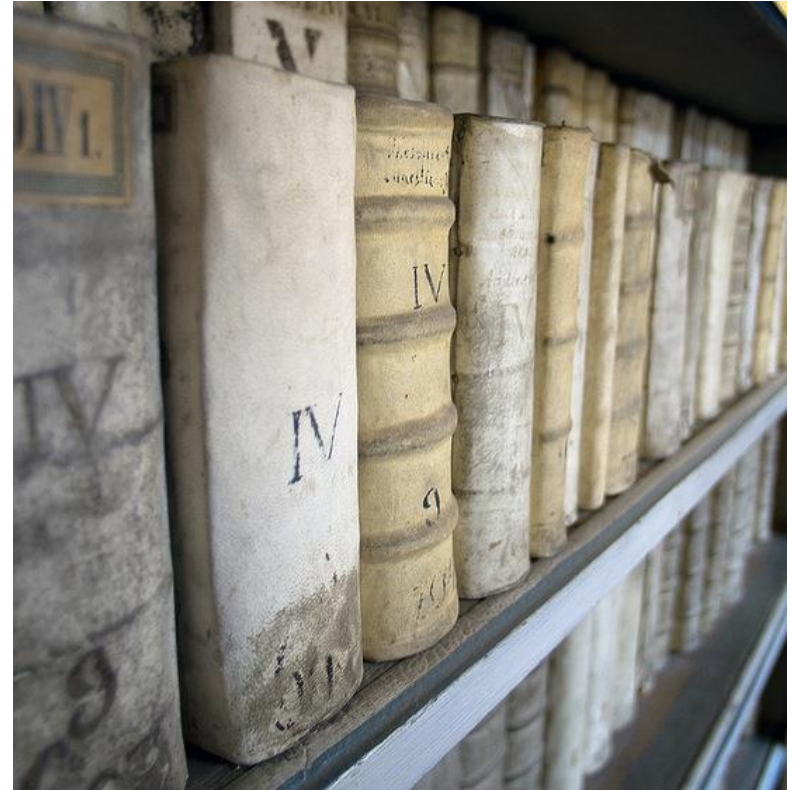
▪ Common features

- Read data from file system, ZIP/JAR archives or classpath
- Preserve directory structure on write for recursive reads



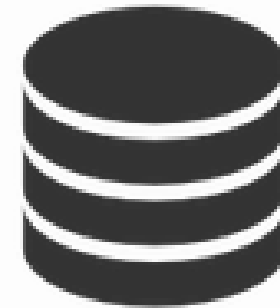
Some currently supported corpora/resources

- British National Corpus
- Wacky Corpora
- TüBa D/Z
- Tiger Corpus
- Digitale Bibliothek
- Brown Corpus
- ACL Anthology Reference Corpus
- ...
- Google Web1T n-grams



Good range of pre-trained models

- Upstream models packaged for convenient deployment and use
- Additional model meta-data
- 90+ models
- 20+ tools
- 15+ languages
- Best supported
 - English (Penn Treebank Tagset, Stanford Dependencies)
 - German (STTS Tagset, Negra/Tiger)



Models/Languages

Various models covering different languages accompany the components.

DKPro Type System Overview

Meta Data

```
DocumentMetaData
<String documentTitle>
<String collectionId>
<String documentId>
<String documentUri>
<String documentBaseUri>
```

Segmentation

Document

Token

LinkingMorpheme

Heading

Compound

Ngram

Paragraph

Split

StopWord

Sentence

CompoundPart

LexMorph

POS

Stem

Lemma

Morpheme

Syntax

Constituent

Dependency

Chunk

Coreference

CoreferenceChain

CoreferenceLink

Semantic Role Labeling

SemanticPredicate

SemanticArgument

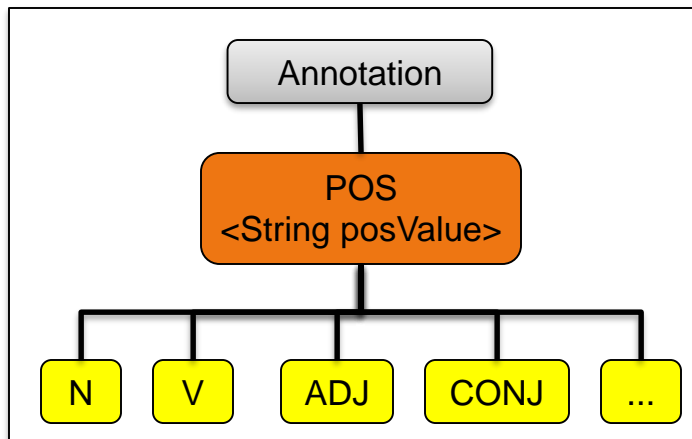
Named Entities

NamedEntity

Location

Person

...etc...

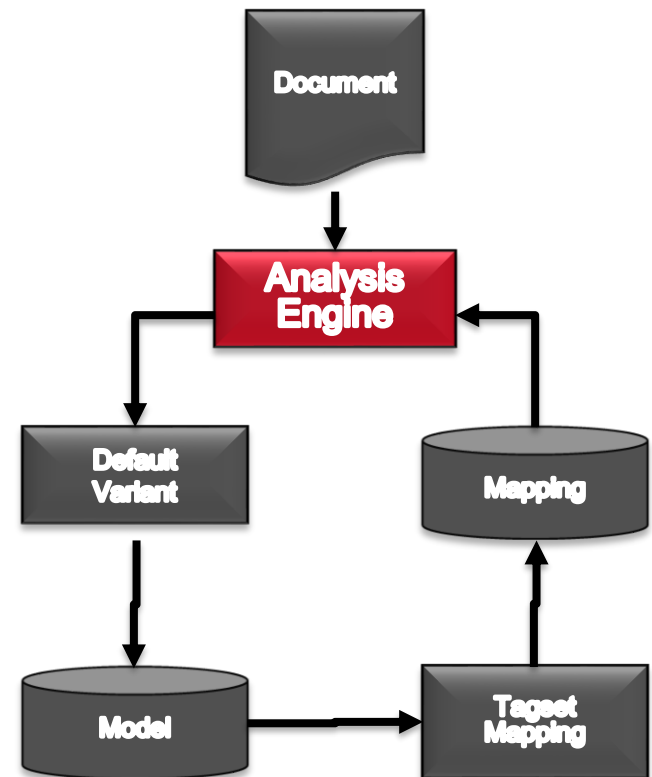


▪ Common parameters

- Model location
- Model encoding
- Model variant
- Mapping location
- Language

▪ Common features

- Model loading based on document language
- Print model tag set to log
- Default variants



Hands-on



TECHNISCHE
UNIVERSITÄT
DARMSTADT