# INCEPTION CORPUS-BASED DATA SCIENCE FROM SCRATCH





Richard Eckart de Castilho, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa, Iryna Gurevych

**Ubiquitous Knowledge Processing Lab (UKP-TUDA) Dept. of Computer Science Technische Universität Darmstadt** 

### Motivation

- Corpus-based data science is seeing rapid adoption in science and industry
- Domain-specific annotated corpora are required in many domains
- **Current corpus processes for manual** corpus annotation do not scale

INCEpTION is an infrastructure-ready human-in-the-loop annotation platform combining machine learning and human expertise for rapid domain adaptation

#### Project Workflow Trained "Crowd" Export / Freeze Subcorporation Identify relevant annotation units Curation Project Setup **Annotation** Define data sources Annotate extracted units Manual comparison Define annotation scheme Automatic aggregation With categories from KB Define knowledge scheme Linking to instances in KB **KB** Population Define schema (classes/properties) Add qualifiers to statements

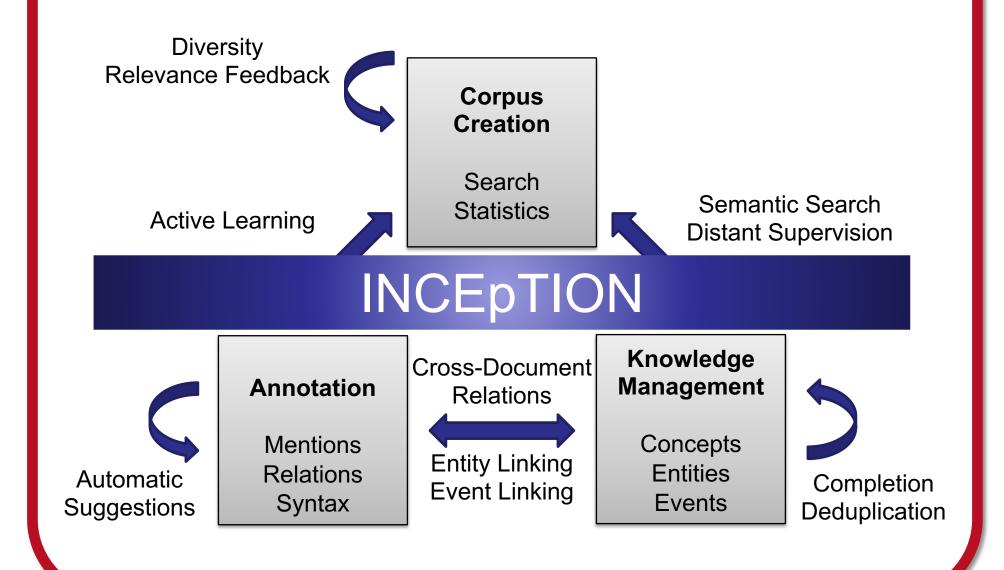
#### **Assisted Annotation** Recommenders **Active Learning** Continually learn from the users actions Aims at reducing the time to learn by asking specific feedback from the user Asynchronous training does not slow Using uncertainty-sampling strategy Automatic evaluation to avoid inaccurate • Compatible with any recommender that predictions (configurable quality threshold) provides a confidence score Built-in recommenders: Dictionary-based, • User can freely switch between active

## INCEpTION Platform

INCEpTION aims to support three functionalities which are commonly required for text annotation projects but typically not available in a single tool: corpus creation; text annotation; knowledge management.

The platform additionally provides assistive features such as machine learning recommenders to help users working on these tasks more efficiently.

Integrating these functionalities into a single comprehensive platform permits addressing tasks typically not found in generic annotation platforms, such as entity linking, knowledge base population, cross-document coreference annotation, etc.



## Sauternes

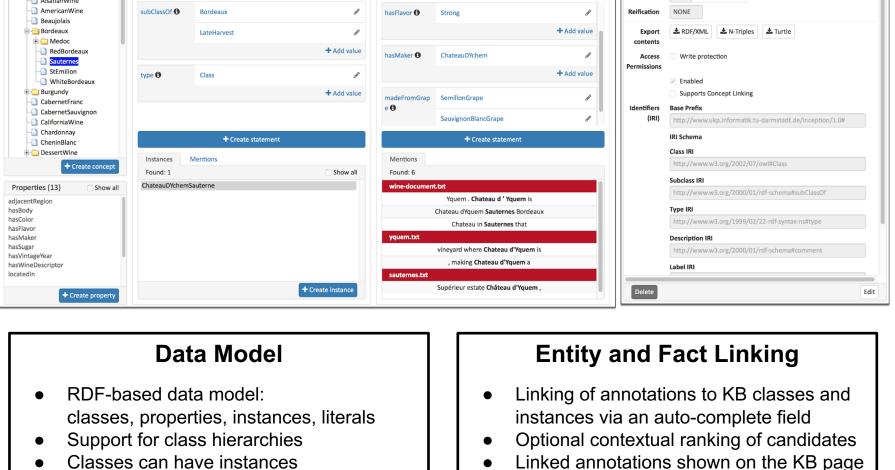
Knowledge Bases

Monitoring

Assign workload Monitor progress

Evaluate quality

Curator



when a class/instance is selected

Special support for linking factual

statements in text documents to

subject/predicate/object triples in the KB

### Extensibility

learning and normal annotation

**Project** 

**Data Formats** 

**Event Logging** 

INCEpTION is a modular architecture providing many extension points where new functionality can be added and existing functionality can be changed - a selection of these is shown below:

Annotation Annotation Importers/Exporters **Editors** Layers **Annotation Editor** Annotation Sidebars Feature types **Annotation Editor** Recommenders Extensions

OpenNLP-based sequence classifier (part-

of-speech, named entities, ...)

The architecture modular architecture is realized using the Spring framework. Dependency injection and events are used to achieve a loose coupling between the modules.

# Open Development

While most annotation tools are built in annotation projects, INCEpTION is an infrastructure software project and is not associated with any single annotation project. Acquiring early adopters and aligning with their use-cases is a key part of our mission.

This motivates our open development philosophy:

- All code is open and publicly available on GitHub under the liberal Apache License 2.0
- All development-related tasks and issues are publicly managed and discussed via GitHub
- Internal and private communication is kept at a minimum

# Interoperability and Integration

Within the NLP and text mining landscape, an annotation platform like INCEpTION only covers a part of the overall text analysis needs. Therefore, it is important that the platform is open and interoperable with external services and resources.

Support for typed properties including

numeric, boolean, KB resources)

Editors for different value types (string,

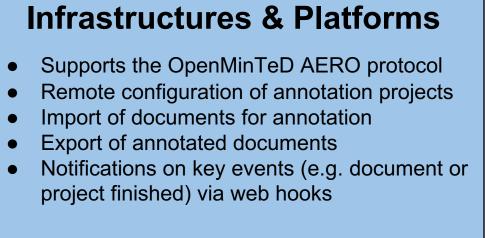
domain and range restrictions

#### **Knowledge Resources** Wikipedia RDF-based knowledge resources are Knowledge bases are accessed via SPARQL DBPedia Configurable schema mapping allows supporting many different knowledge resources (RDFS, OWL, SKOS, ...) Support for entity linking if knowledge base Custom KB has full-text search capabilities **Annotation Services** DKPro TC Connect external NLP services and machine

- learning tools to support the user during the annotation process Programming language independent HTTP-
- Protocol supports re-training the external tools UIMA CAS XMI data format; supported in Python via DKPro PyCAS; in Java via Apache UIMA

System-level integration

Integration goes beyond interoperability. E.g. when an external text mining platform wants to delegate annotation to the INCEpTION platform, it needs to be able to automatically set up annotation projects, import data, monitor the ongoing annotations and finally retrieve the annotated data for further use.



# INCEPTION

#### **Authentication**

- Support for external authentication mechanisms via HTTP headers Allows delegating authentication to a reverse proxy (e.g. Apache HTTPD) and using any of its authentication mechanisms, e.g. LDAP,
- SAML2/Shibboleth Enables single-sign-on scenarios

Legacy **CLARIN TCF** CoNLL-U **CoNLL** WebAnno TSV **Data Formats UIMA CAS** Binary+XMI • Supports a range of different formats for

importing and exporting annotated text corpora

interoperability with the CLARIN-D WebLicht

including the import of entire annotation projects

Full compatibility with CLARIN's WebAnno

• UIMA CAS data format enables interoperability

Support of the TCF format enables

annotation services

#### **Annotation Schemata**

Flexible type system configuration. Use UIMA type system descriptor to autoconfigure an annotation project.

with NLP pipelines like DKPro Core

DKPro Core types pre-configured (token, sentence, POS, lemma, morphological features, dependencies, coreference chains...)

Data-level interoperability





Python-

based

machine



Download INCEpTION at https://inception-project.github.io

This work has received funding from the German Research Foundation under the grants No. EC 503/1-1 and GU 798/21-1 (INCEpTION) and as part of the RTG Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1. We thank Wei Ding, Peter Jiang, Marcel de Boer and Michael Bugert for their valuable contributions.



Text