INCEpTION
Corpus-based Data Science from Scratch

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

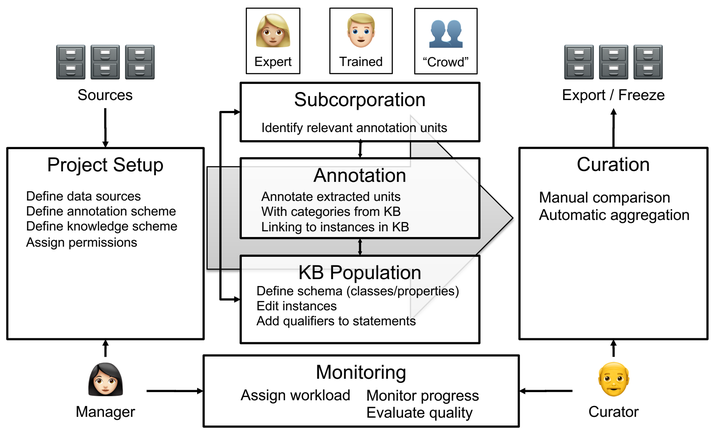
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Dept. of Computer Science
Technische Universität Darmstadt

Motivation

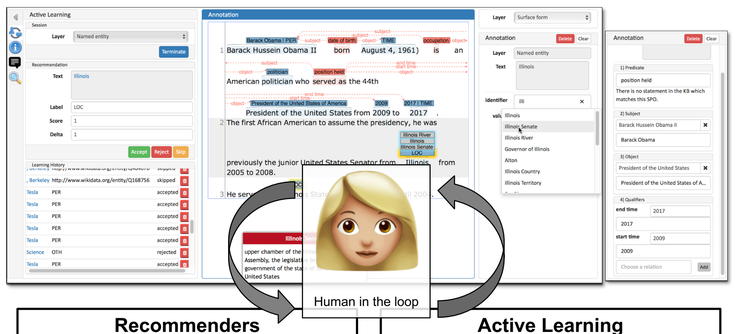
- **Corpus-based data science is seeing rapid adoption in science and industry**
- **Domain-specific annotated corpora are required in many domains**
- **Current corpus processes for manual corpus annotation do not scale**

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

Project Workflow



Assisted Annotation



Recommenders

- Continually learn from the users actions
- Asynchronous training does not slow down user interface
- Automatic evaluation to avoid inaccurate predictions (configurable quality threshold)
- Built-in recommenders: Dictionary-based, OpenNLP-based sequence classifier (part-of-speech, named entities, …)

Active Learning

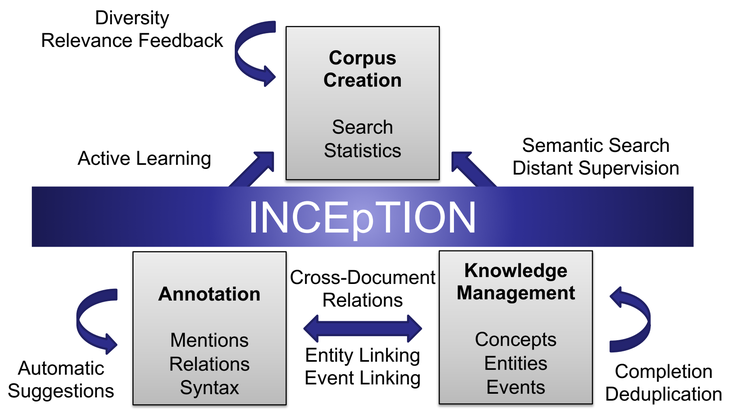
- Aims at reducing the time to learn by asking specific feedback from the user
- Using uncertainty-sampling strategy
- Compatible with any recommender that provides a confidence score
- User can freely switch between active learning and normal annotation

INCEpTION Platform

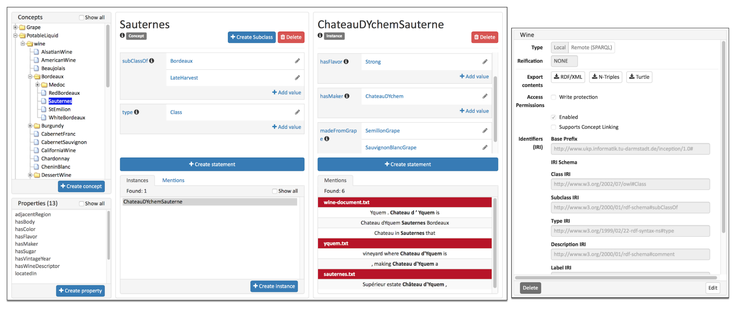
INCEpTION aims to support three functionalities which are commonly required for text annotation projects but typically not available in a single tool: corpus creation; text annotation; knowledge management.

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.



Knowledge Bases



Data Model

- RDF-based data model: classes, properties, instances, literals
- Support for class hierarchies
- Classes can have instances
- Support for typed properties including domain and range restrictions
- Editors for different value types (string, numeric, boolean, KB resources)

Entity and Fact Linking

- Linking of annotations to KB classes and instances via an auto-complete field
- Optional contextual ranking of candidates
- Linked annotations shown on the KB page when a class/instance is selected
- Special support for linking factual statements in text documents to subject/predicate/object triples in the KB

Extensibility

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 Editors	Annotation Layers	Project Importers/Exporters
Annotation Editor Sidebars	Annotation Feature types	Data Formats
Annotation Editor Extensions	Recommenders	Event Logging

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.

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.

Deutsche Forschungsgemeinschaft
AIPHES — Adaptive Preparation of Information from Heterogeneous Sources

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.