BEYOND WEBANNO:
THE INCEpTION TEXT ANNOTATION PLATFORM
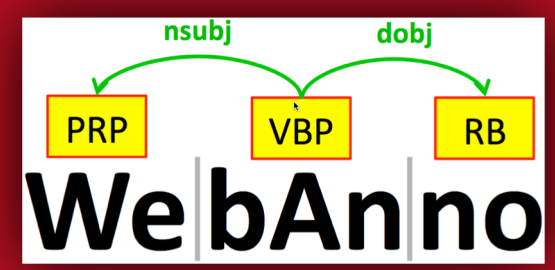
Richard Eckart de Castilho, Jan-Christoph Klie, Ute Winchenbach, Iryna Gurevych

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

WebAnno → INCEpTION

- WebAnno originated in CLARIN-D and is deployed in several national CLARIN infrastructures, used by various research groups as well as individual researchers
- INCEpTION takes the technology to a new level and provides a human-in-the-loop annotation platform combining machine learning and human expertise for rapid domain adaptation
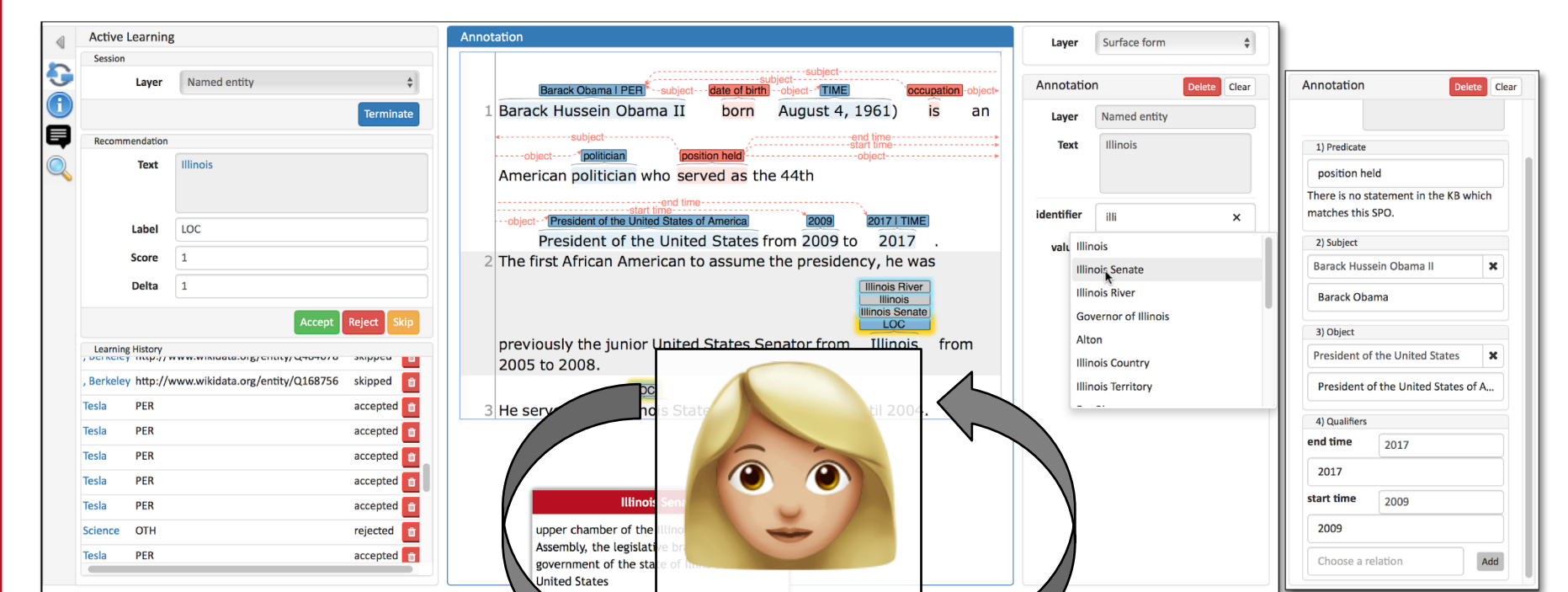
Upgrade today!

Migration from WebAnno to INCEpTION is straightforward!

Project Workflow
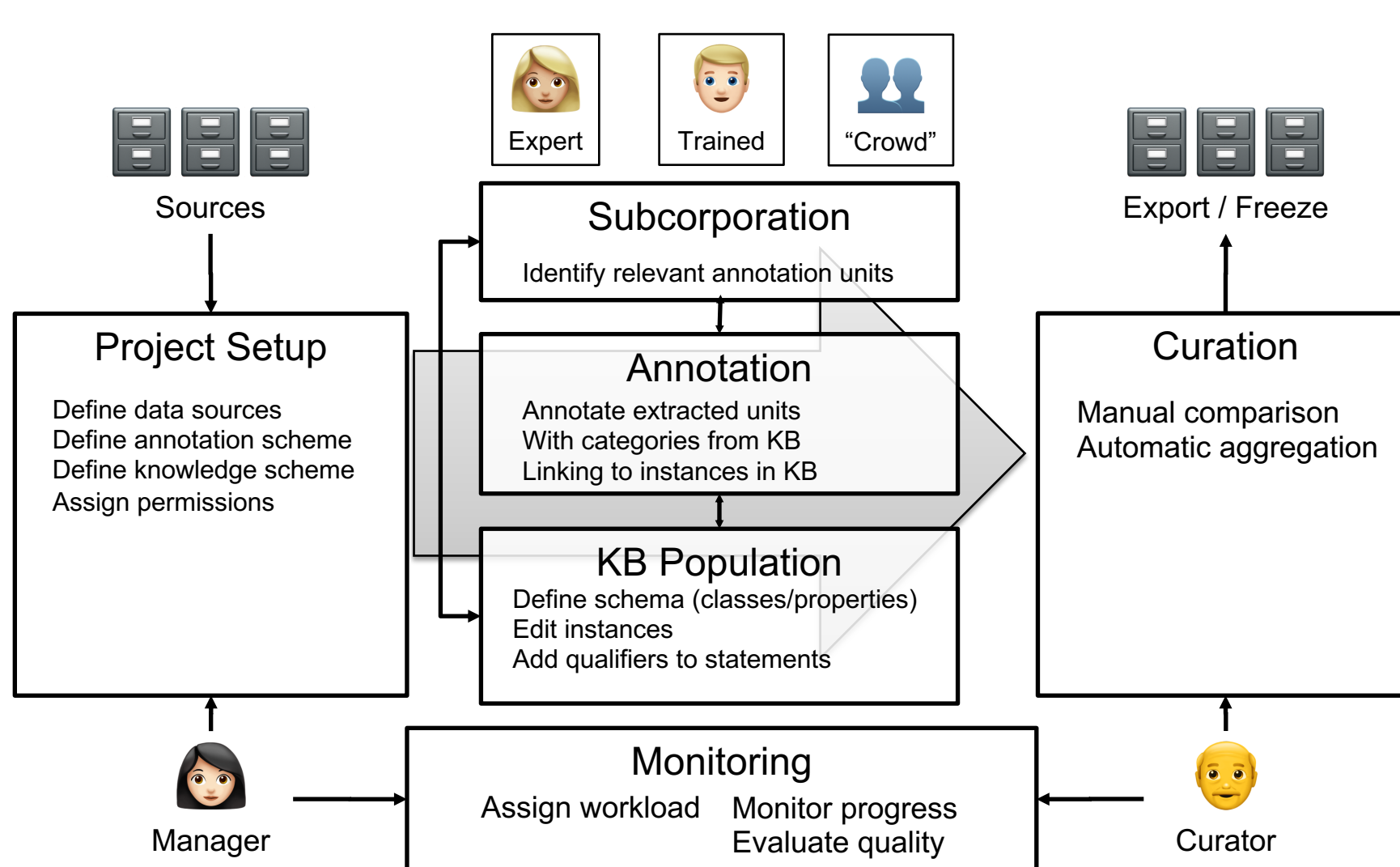
Assisted Annotation

Human in the loop

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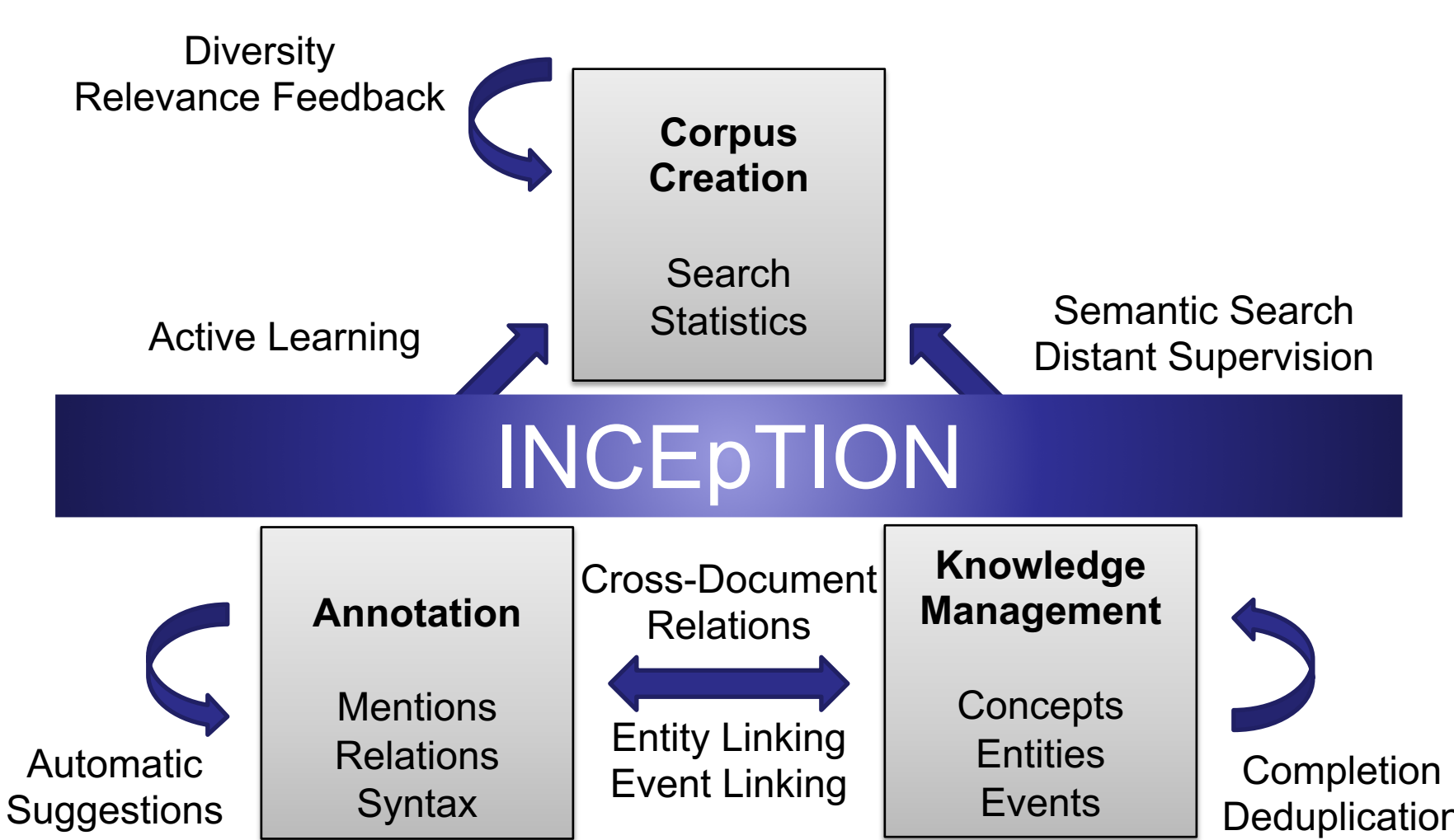
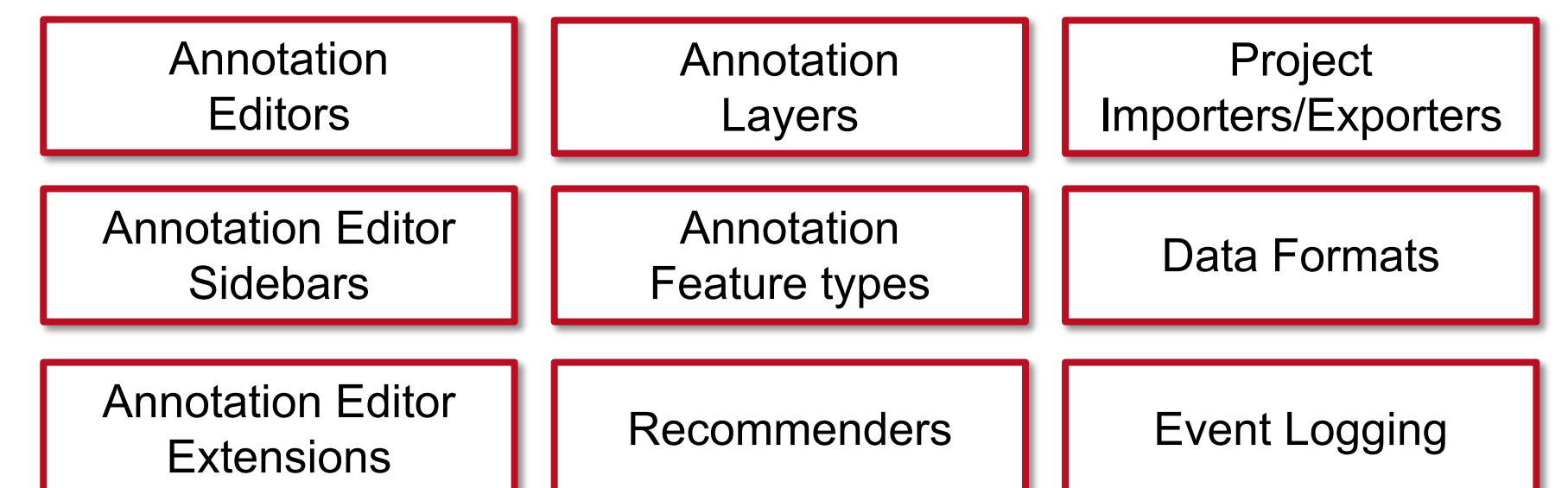## Knowledge Bases
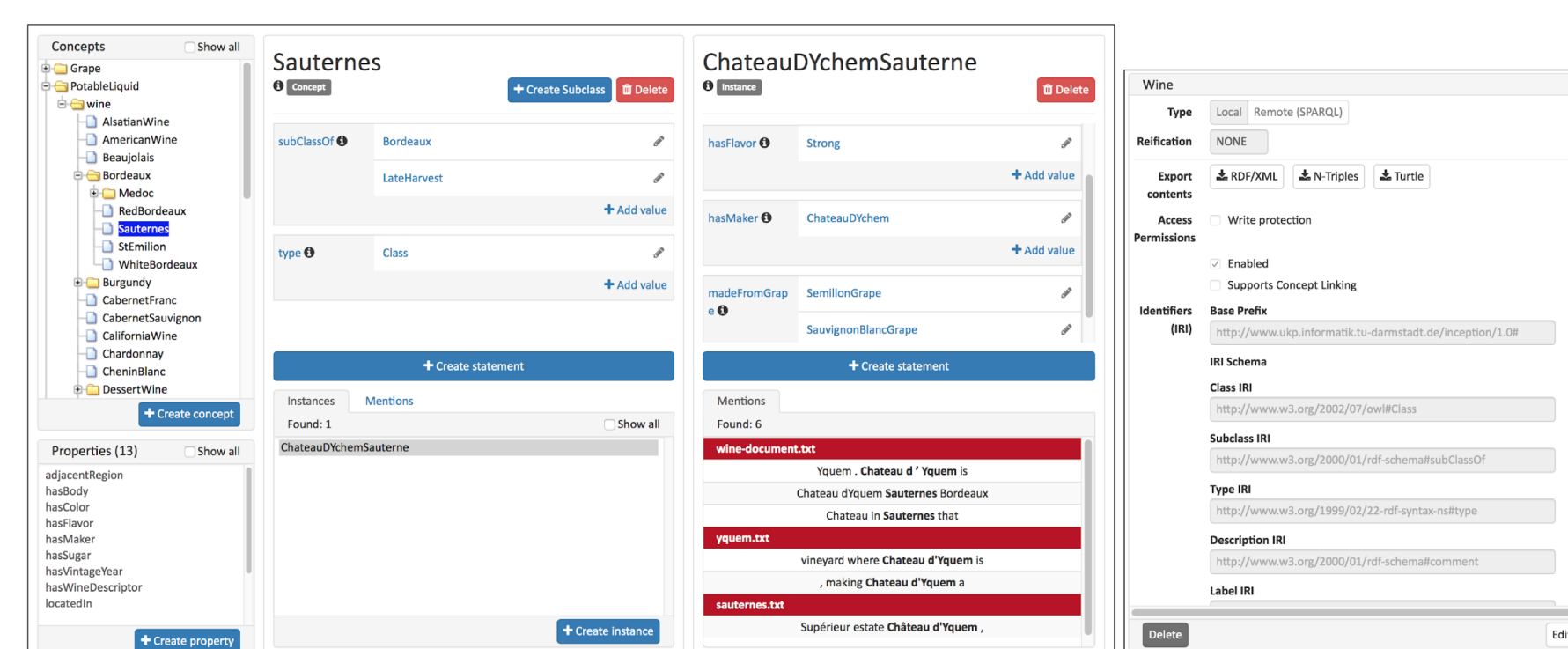
Data Model
- RDF-based data model: classes, properties, instances, literals
- Support for class hierarchies
- Classes can have instances
- Support for typed properties including domain and range restrictions
- Editors for different value types (string, numeric, boolean, KB resources)

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:

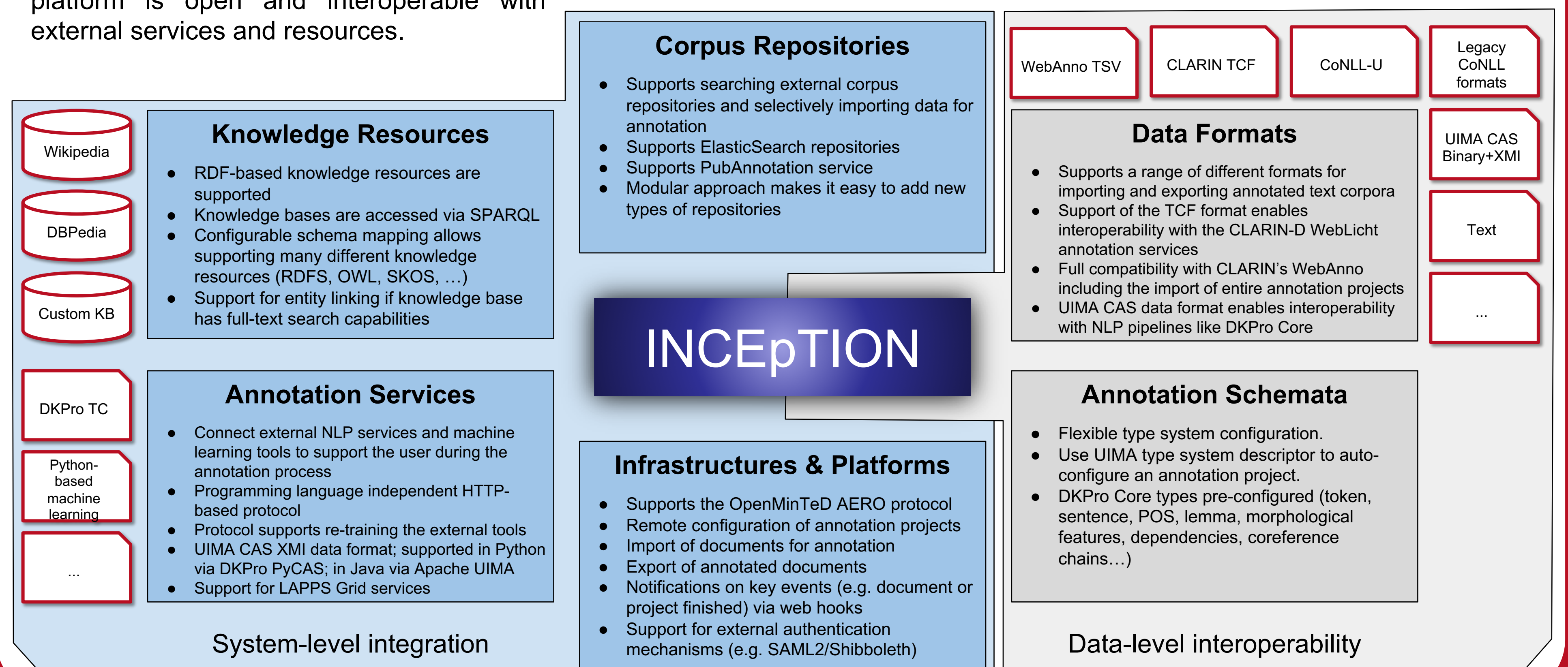|		
Annotation Editors	Annotation Layers	Project Importers/Exporters
Annotation Editor Sidebars	Annotation Feature types	Data Formats
Annotation Editor Extensions	Recommenders	Event Logging

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

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.

Knowledge Resources
- RDF-based knowledge resources are supported
- Knowledge bases are accessed via SPARQL
- Configurable schema mapping allows supporting many different knowledge resources (RDFS, OWL, SKOS, …)
- Support for entity linking if knowledge base has full-text search capabilities

Wikipedia, DBPedia, Custom KB

Corpus Repositories
- Supports searching external corpus repositories and selectively importing data for annotation
- Supports ElasticSearch repositories
- Supports PubAnnotation service
- Modular approach makes it easy to add new types of repositories

WebAnno TSV, CLARIN TCF, CoNLL-U, Legacy CoNLL formats

Data Formats
- Supports a range of different formats for importing and exporting annotated text corpora
- Support of the TCF format enables interoperability with the CLARIN-D WebLicht annotation services
- Full compatibility with CLARIN's WebAnno including the import of entire annotation projects
- UIMA CAS data format enables interoperability with NLP pipelines like DKPro Core

UIMA CAS Binary+XMI, Text

Annotation Services
- Connect external NLP services and machine learning tools to support the user during the annotation process
- Programming language independent HTTP-based protocol
- Protocol supports re-training the external tools
- UIMA CAS XMI data format; supported in Python via DKPro PyCAS; in Java via Apache UIMA
- Support for LAPPS Grid services

DKPro TC, Python-based machine learning

Infrastructures & Platforms
- Supports the OpenMinTeD AERO protocol
- Remote configuration of annotation projects
- Import of documents for annotation
- Export of annotated documents
- Notifications on key events (e.g. document or project finished) via web hooks
- Support for external authentication mechanisms (e.g. SAML2/Shibboleth)

Annotation Schemata
- Flexible type system configuration
- Use UIMA type system descriptor to auto-configure an annotation project
- DKPro Core types pre-configured (token, sentence, POS, lemma, morphological features, dependencies, coreference chains…)

System-level integration

Data-level interoperability

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