

DKPro Core and INCEpTION

Modular, interoperable, reusable TDM tools for the community

Richard Eckart de Castilho



This presentation includes protected logos, brand names, and trade marks owned by their respective owners. These are used as visual citations of the respective product and brand names. The usage of these media resources does not indicate any endorsement of DKPro Core, INCEpTION, Technische Universität Darmstadt, VisaTM, the speaker or any other aspects of this presentation and its content. Excluding such logos, the slides are provided under the Creative Commons CC-BY-SA 4.0 license.

Photo: Poppies in Southern France, 2009 - Richard Eckart de Castilho, private archive

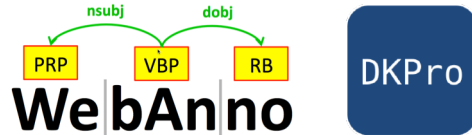


Richard Eckart de Castilho

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

- PostDoc @ UKP
- Open source guy
- Java developer
- INCEpTION PI
- Apache UIMA developer
- DKPro and WebAnno person
- NLP software infrastructure researcher

INCEpTION





Ubiquitous knowledge Processing LAB

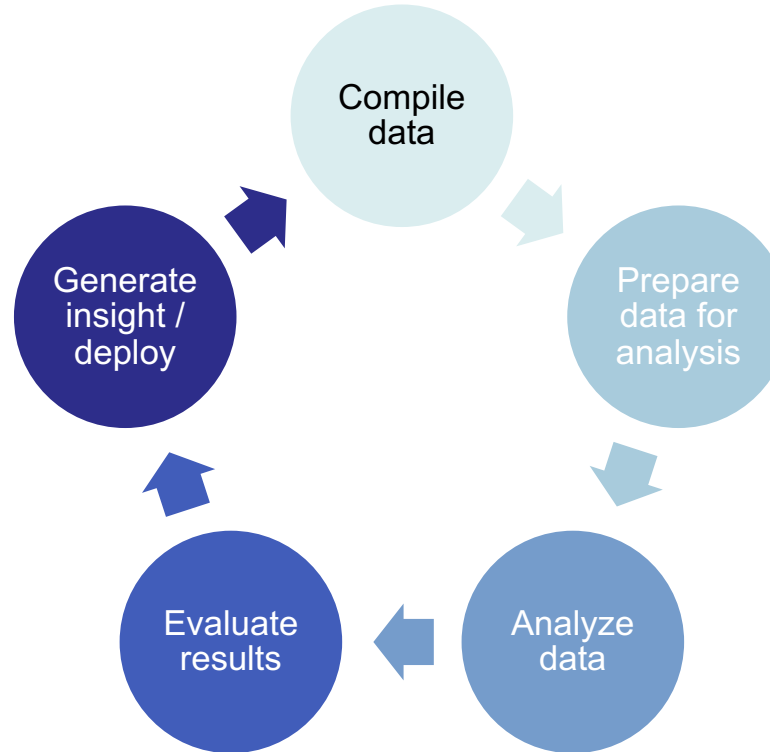
Prof. Dr. Iryna Gurevych
Technische Universität Darmstadt

- Argumentation Mining
- Language Technology for Digital Humanities
- Lexical-Semantic Resources & Algorithms
- Text Mining & Analytics
- Writing Assistance and Language Learning

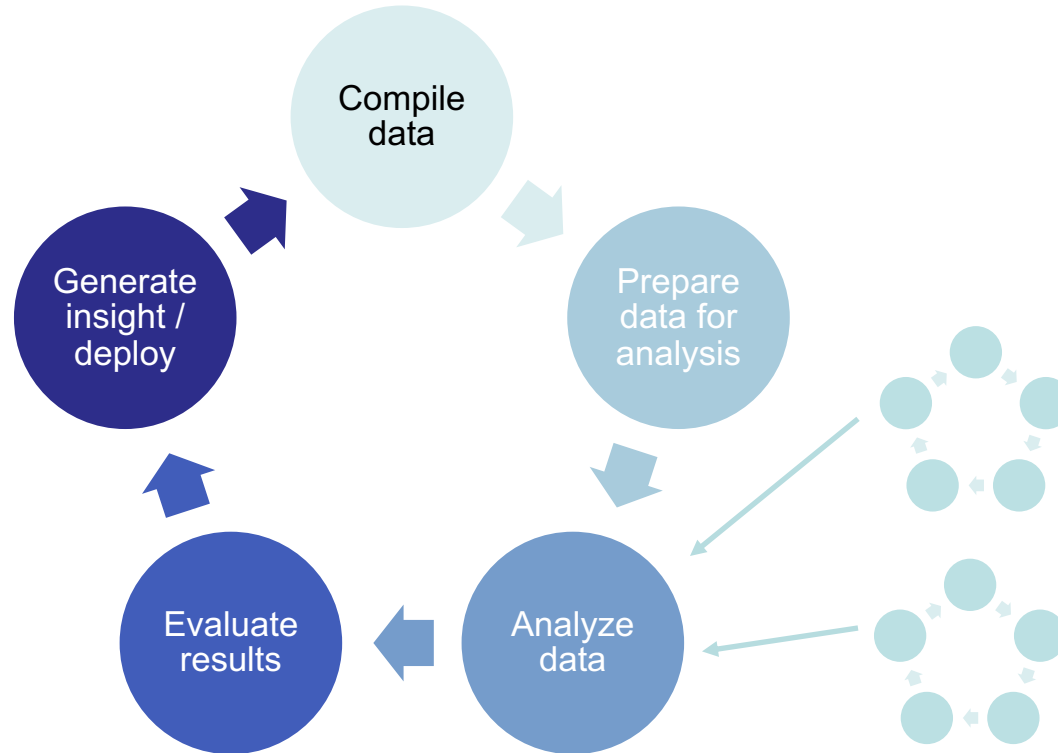
 @UKPLab

<http://www.ukp.tu-darmstadt.de>

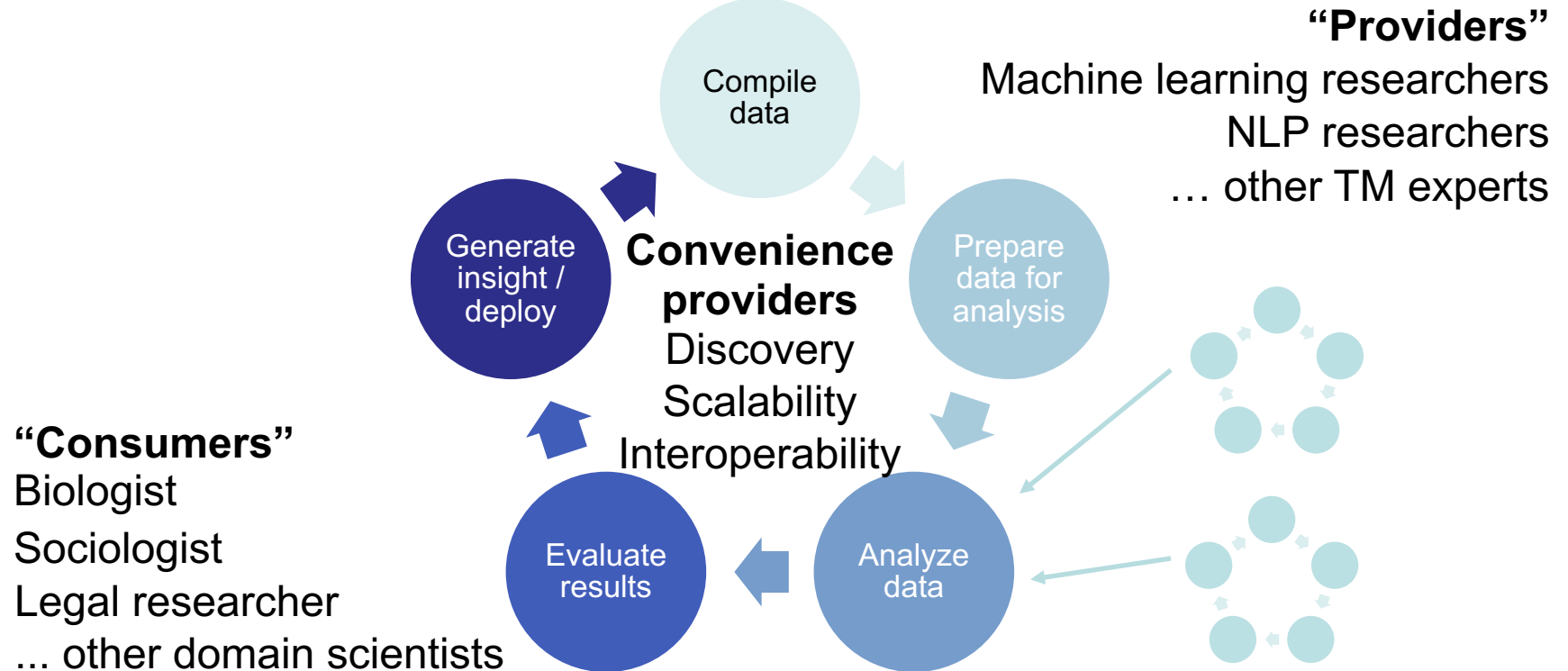
The Text Mining Process



The Text Mining Process



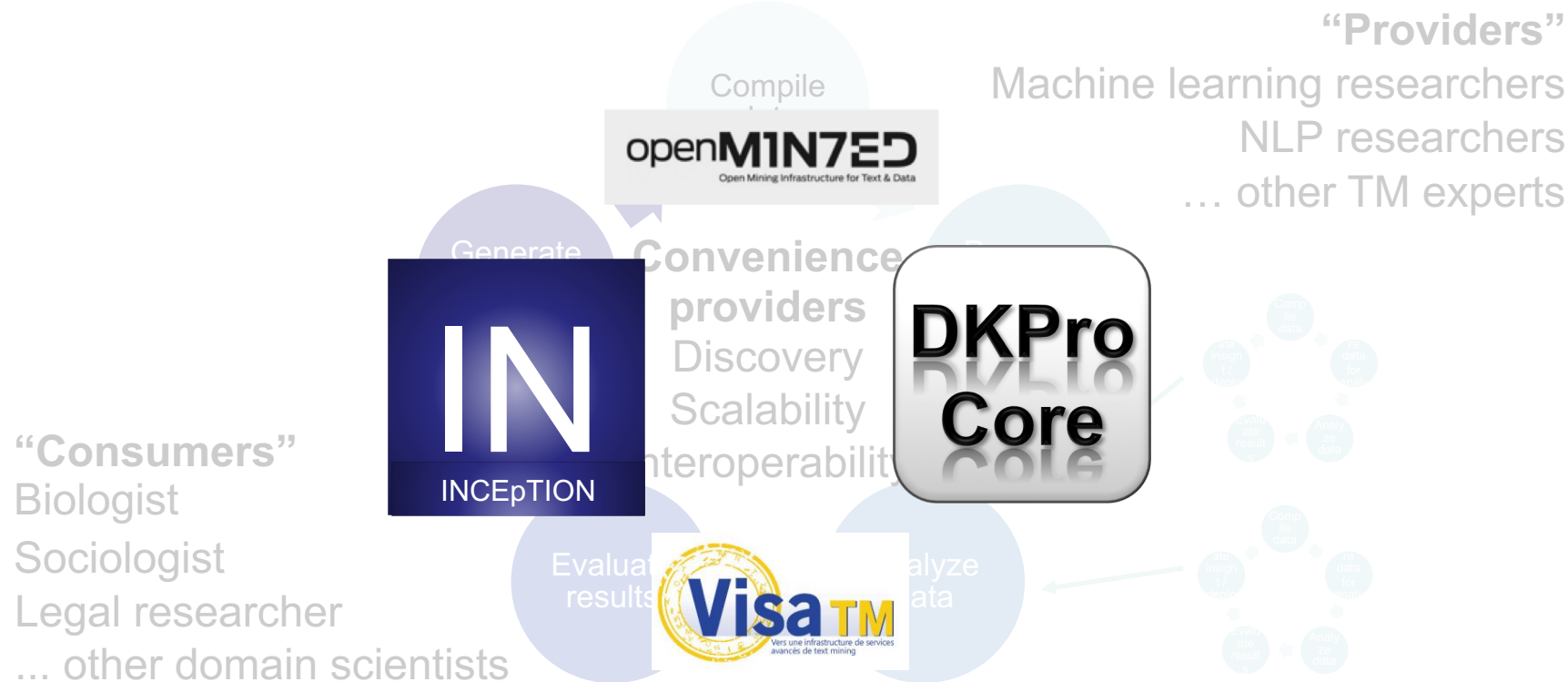
Who are the stakeholders?



Who are the stakeholders?



TECHNISCHE
UNIVERSITÄT
DARMSTADT



UIMA-based linguistic preprocessing

DKPro Core



- Based on Apache UIMA
 - Natural Language Processing
 - Pre-processing for ML / AI
 - Mix & match components
 - Convert between formats
 - Train models
 - Evaluate
-
- Experimental pipelines
 - Embed in applications
 - Ready to run on server/cluster

DKPRO CORE DOWNLOADS DOCUMENTATION ISSUES SOURCE CONTACT ABOUT

DKPro Core - Welcome

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](#)

Latest release: 2.0.0 (2019-06-09)
maven central 2.0.0

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.

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.

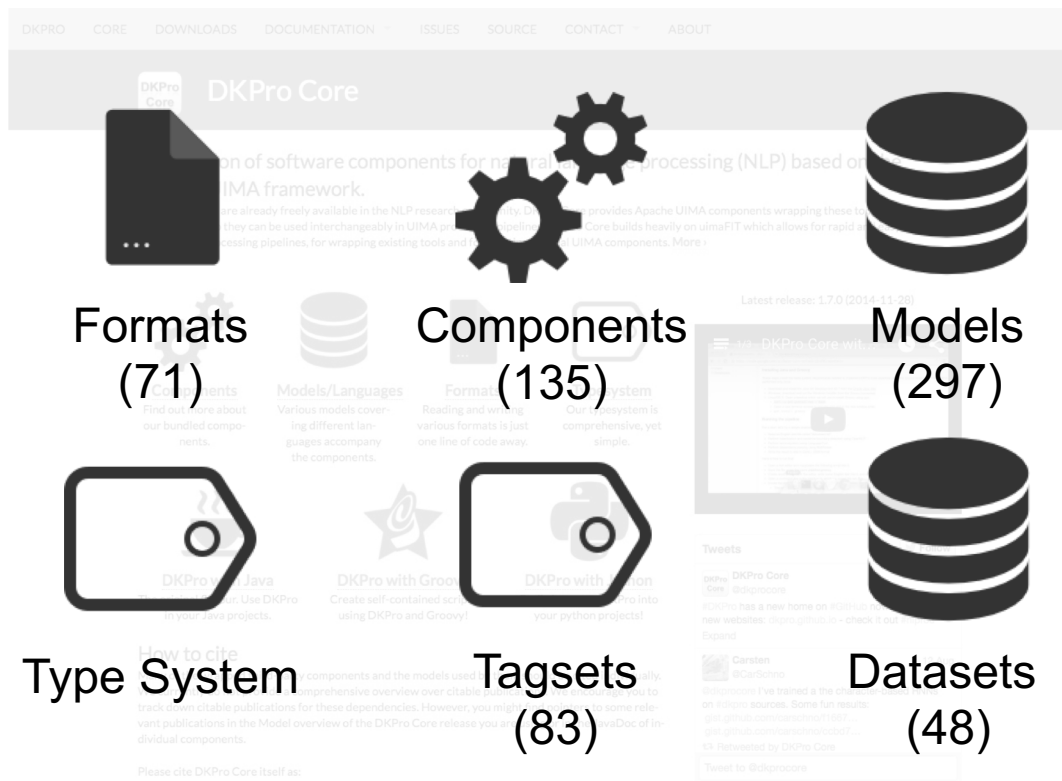
Follow us on Twitter

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

@dkprocore

DKPro Core

Building blocks



Chunker
Coreference resolver
Language Identifier
Lemmatizer
Morphological analyzer
Named entity recognizer
Parser
Part-of-speech tagger
Phonetic transcriptor
Segmenter
Semantic role labeler

Building pipelines with DKPro Core is easy



```
@Grab('org.dkpro.core:dkpro-core-opennlp-asl:2.0.0')
@Grab('org.dkpro.core:dkpro-core-corenlp-gpl:2.0.0')
@Grab('org.dkpro.core:dkpro-core-lingpipe-gpl:2.0.0')
```

Fetches all required
dependencies
No manual installation
Input

```
// ... 6 extra lines importing the necessary function calls ...
```

```
def document = createText('''\
  In this study, we investigated the roles of serum amyloid A (SAA) in T helper 17 \
  (Th17)-related cytokine induction in rheumatoid arthritis (RA) synoviocytes. \
  Synoviocytes isolated from rheumatoid arthritis (RA) patients were stimulated \
  with recombinant SAA and IL-23 expression was investigated using reverse \
  transcriptase-polymerase chain reaction and Western blot. \
  ''', "en");
```

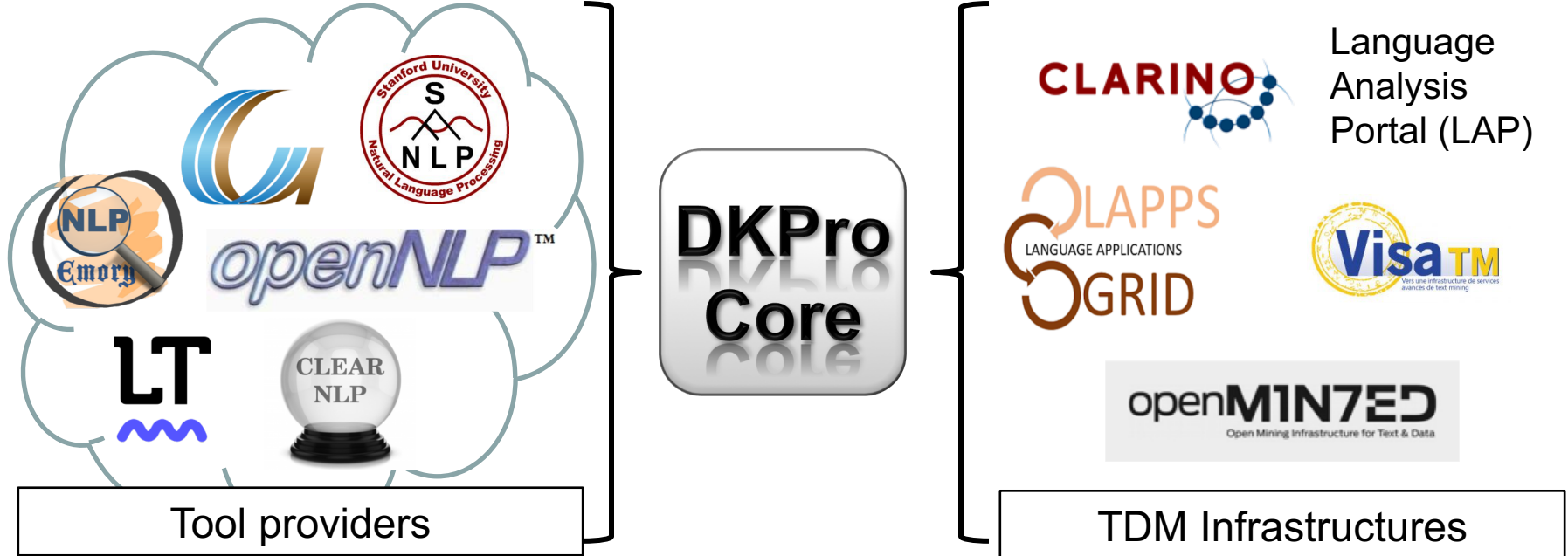
```
runPipeline(document,
  createEngineDescription(org.dkpro.core.opennlp.OpenNlpSegmenter),
  createEngineDescription(org.dkpro.core.corenlp.CoreNlpPosTagger),
  createEngineDescription(org.dkpro.core.lingpipe.LingPipeNamedEntityRecognizer,
    'modelVariant', 'bio-genia'));
```

Analytics pipeline.
Language-specific
resources fetched
automatically

```
select(document, Token).each { println "${it.coveredText} ${it.pos.posValue}" }
select(document, NamedEntity).each { println "${it.value}: [${it.coveredText}]" }
```

Output

Bridging the gap between text mining tools and infrastructures



Logos are property of their respective owners: LanguageTool, Apache OpenNLP, EmoryNLP, ClearNLP, Stanford CoreNLP, Cognitive Computation Group @ U. Penn.

Logos are property of their respective owners: CLARINO (Norway), Language Application Grid (USA), OpenMinTeD (EU), VisaTM (France)

Tasks



Corpus Creation

Search
Statistics

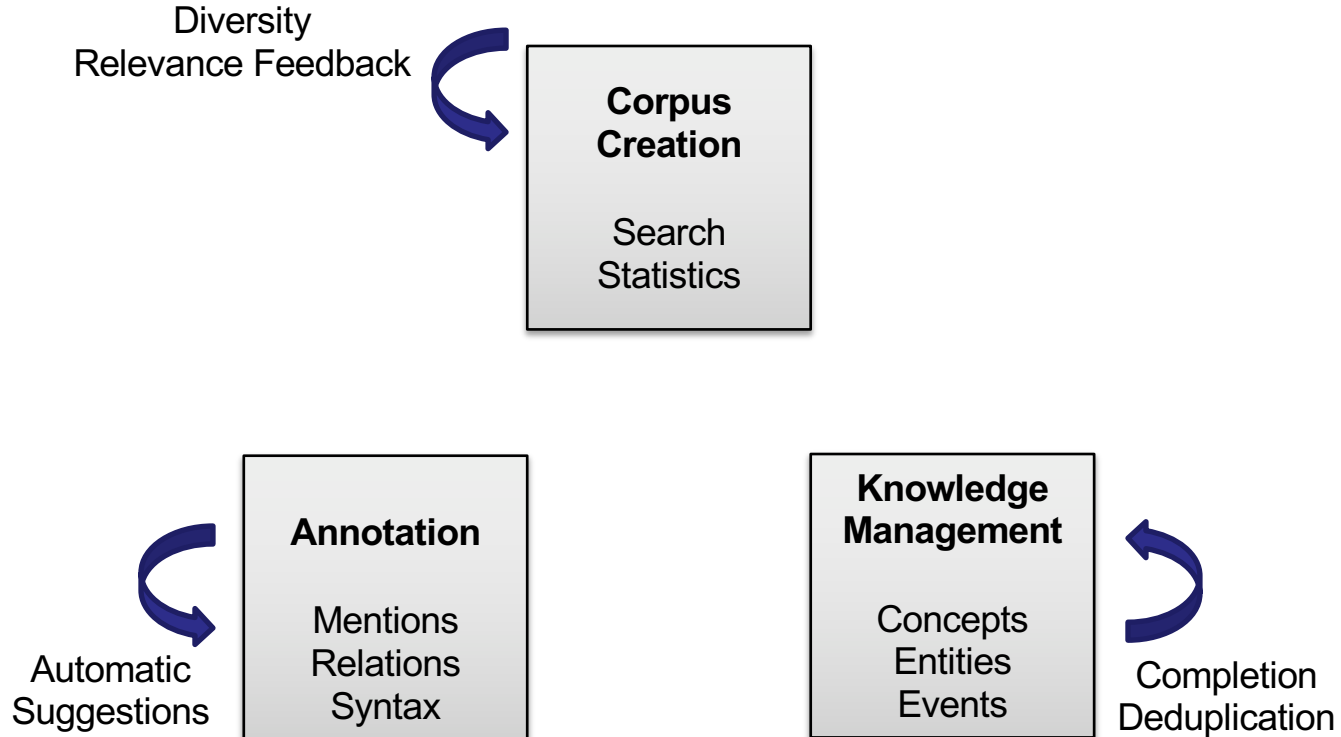
Annotation

Mentions
Relations
Syntax

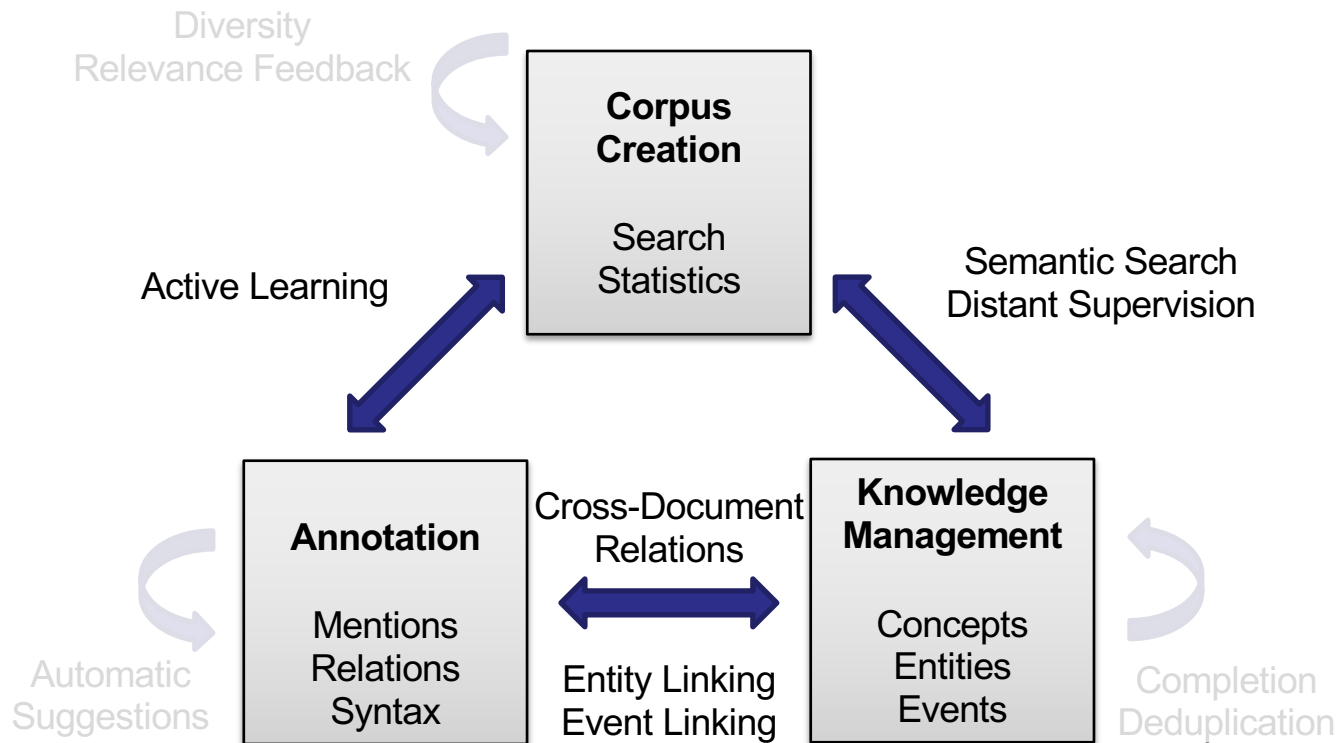
Knowledge Management

Concepts
Entities
Events

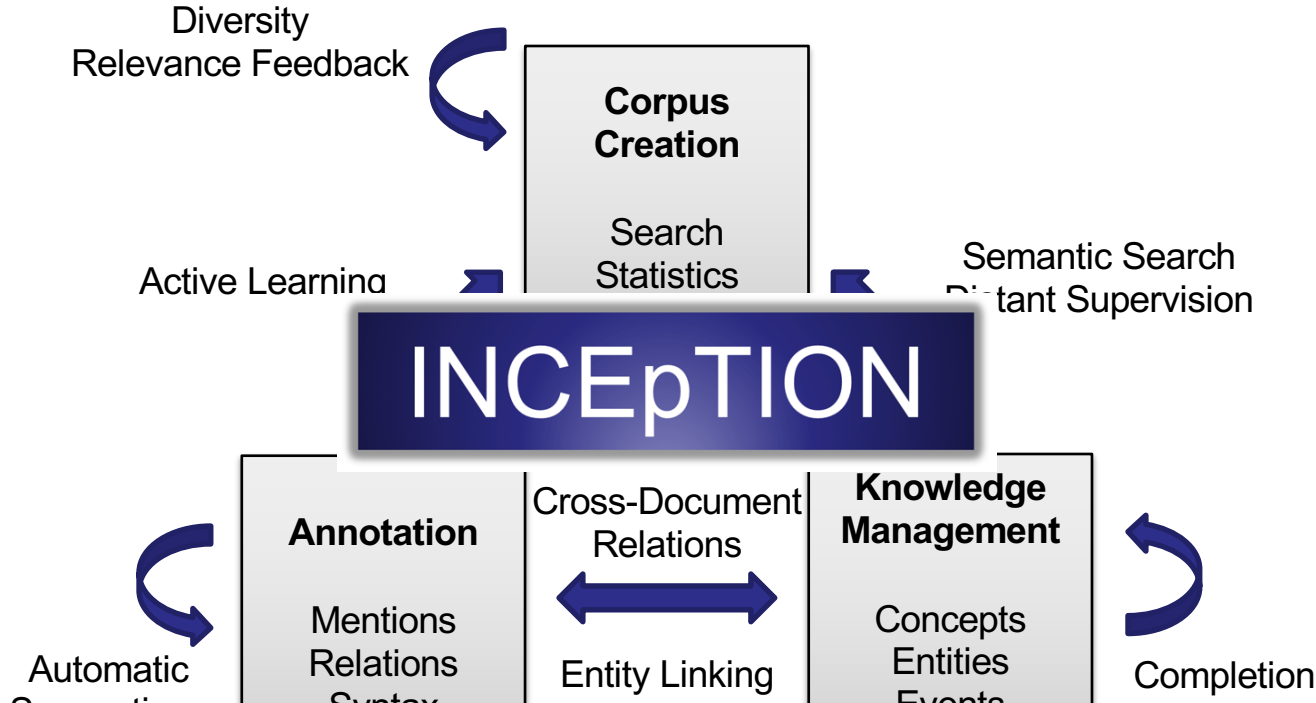
Enhancements for Individual Tasks



Interactions between Tasks



Integrated User Experience / Modular Software Platform



<https://inception-project.github.io>



#tap_inception

Annotation

Layer Surface form

Annotation Delete Clear

Layer Named entity

Text Illinois

Identifier illi

value

- Illinois
- Illinois Senate
- Illinois River
- Governor of Illinois
- Alton
- Illinois Country
- Illinois Territory

Illinois Senate

upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States

Human-in-the-loop Text Analysis



TECHNISCHE
UNIVERSITÄT
DARMSTADT

INCEpTION Projects Dashboard

Help Administration rec Log out (automatically in 29 min)

Recommender

Settings

Max. suggestions: 3

Show all: ☐

Retrain Apply

Learning curve

Recommenders

[Named entity@value] Multi-Token Sequence Classifier (OpenNLP NER) active

F1: 0.415 Acc.: 0.714

Prec.: 0.5 Rec.: 0.355

Accept all

[Named entity@value] String Matcher active

F1: 0.23 Acc.: 0.133

Prec.: 0.667 Rec.: 0.139

Accept all

rec: REC Recommendation Playground (NER)/Wikipedia-Obama.txt Showing 1-10 of 28 sentences [document 1 of 1]

1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017.

2 The first African American to a

3 He served in the Illinois State Senate from 1997 until 2004.

4 Obama was born in 1961 in Honolulu, Hawaii, two years after the territory was admitted to the Union as the 50th state.

5 Raised largely in Hawaii, Obama also spent one year of his childhood in Washington State and four years in Indonesia.

6 After graduating from Columbia University in New York City in 1983, he worked as a community organizer in Chicago.

7 In 1988 Obama enrolled in Harvard Law School, where he was the first black president of the Harvard Law Review.

8 After graduation, he became a civil rights attorney and professor, and taught

Layer: Named entity

Annotation

No annotation selected

Named entity

"Illinois State Senate"

Note: [Named entity@value] Multi-Token Sequence Classifier (OpenNLP NER)

Confidence: 0.72

ID: recommendationEditorExtension:3-38142

Technische Universität Darmstadt -- Computer Science Department -- INCEpTION -- 0.13.0 (2019-11-11 19:02:15, build c666f4757c37b57a20cfb8593ddbba9c6c02b1ea)

INCEpTION in the TDM Landscape



Integrated Machine Learning



HTTP API
for custom
machine
learning



Ext. processes

OpenMinTeD
AERO API
for project
management

Text Mining Services



Knowledge Management



INCEpTION

Knowledge Services



Document Repositories



Foundations



Unstructured
Information Management
Architecture
An Apache Project.

Repository Services



Logos are property of their respective owners. They are used to represent which products and standards INCEpTION is compatible with.

Open Development Strategy



Open communication and development planning

Provide ideas & prototypes
Prioritize based on synergies and user feedback

Open code on a contributor-friendly platform

Everything accessible
Contribution guidelines
Minimal hurdles

Liberal Licensing

Minimal restrictions
Maximal impact
Treat all equally:
researchers, students,
businesses, hobbyists

Quality assurance and testing

Be always ready to release
Control risks
Minimize overhead

Friendly support and helpful tutorials

Stay in contact with the users
Be friendly and inviting
Take up user feedback

Public software distribution and easy deployment

Minimize hurdles for adoption
Release often
Minimize update overhead

Essential Tools

Open communication and
development planning



Open code on a
contributor-friendly
platform



Liberal Licensing



Quality assurance
and testing



Friendly support
and helpful tutorials



Public software distribution
and easy deployment



Know-how and Human Capital



Open communication and
collaboration

Open code on a
platform

Liberal Licensing

Code is infrastructure

Technical know-how ➡ maintenance & development

Social networking ➡ use-cases & community building

Requires staff continuity

Requires cross-institutional collaboration and knowledge redundancy

**Software managers and engineers
that maintain and develop the code are also part of the infrastructure (?)**



Thank you!

Thanks to the VisaTM and
OpenMinTeD teams!

The INCEpTION project is joint work with
Prof. Dr. Iryna Gurevych
Jan-Christoph Klie
Ute Winchenbach
...and our users and contributors

The DKPro family and the DKPro Core project are joint work
with all the users and open source contributors

INCEpTION

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft

