

# Anonymization of court decisions in Germany

## An essential requirement for E-Justice

**Michael Keuchen**, [michael.keuchen@fau.de](mailto:michael.keuchen@fau.de)  
Scientific assistant at the Law School of the FAU



FRIEDRICH-ALEXANDER  
UNIVERSITÄT  
ERLANGEN-NÜRNBERG

FACHBEREICH  
RECHTSWISSENSCHAFT



LEHRSTUHL FÜR  
KORPUS- UND  
COMPUTERLINGUISTIK

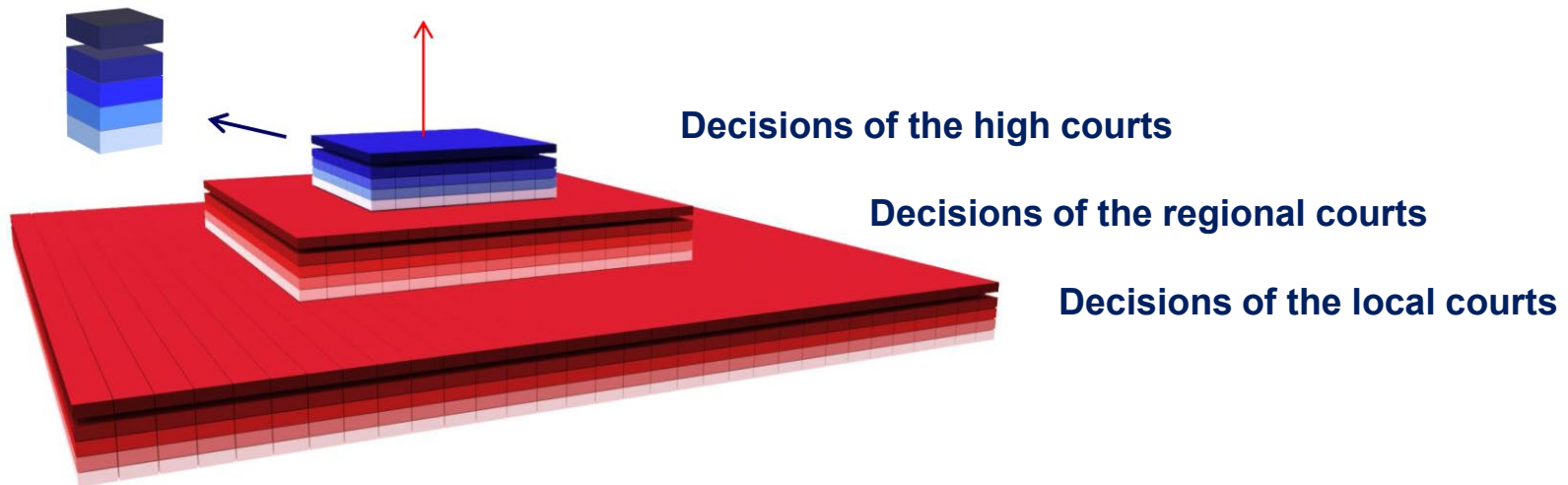
## Why anonymization of court decisions?

Protection of personal identifiable information and legitimate interests.

Publication of court decisions

Decisions as training data for legal tech applications

# Publication practice in Germany



© Axel Adrian

- But only about 2 % of all court decisions are published.
- Most of these decisions originated from courts of last instance and high courts.
- First-instance decisions regarding questions of facts are underrepresented.
- Inconsistency in the features that need to be anonymized and in the used anonymization techniques.

## Obligation to publish court decisions

- There are only a **few statutory** publication obligations in Germany.
- However, a general obligation to publish jurisprudence has been developed by **case law**. The obligation is directly **derived from the constitution**, namely the principle of democracy and the rule of law.
- The publication obligation exists regardless of the court instance, the right to appeal and an objective interest in the decision.
- The **protection of overriding interests** of the parties involved must be ensured. Anonymization is required.

## Anonymity in the GDPR

- Focusing on the protection of natural persons and their individual rights. Legal persons are not included (recital 14 second sentence GDPR).
  - In the german statistics law it is prohibited to link official (anonymized) statistical datasets (e.g. judicial statistics) for deanonymization. Court decisions can be seen as individual linkable datasets. Thus the decisions entail a higher risk of deanonymization than statistical datasets. Therefore an even stricter standard needs to be applied.
  - Risks of deanonymization are still insufficiently researched. Empirical data on effort, time and additional knowledge is required.
- **Rethinking of anonymity is necessary.**

# Which features need to be anonymized?

## **Direct** identifiers:

- Names (natural and legal persons)
- Addresses
- Dates of birth
- ...

## **Indirect** identifiers:

- Profession details
- Academic titles
- Health data
- Descriptive informations about local conditions or companies
- Unique features (e.g. the only red house in a small village)

→ **Risk of cross-referencing**

# Research project on the automatic anonymization of court decisions

## Goal:

Evaluation of the legal and technical issues concerning the ability to automatically anonymize decisions, in particular by using corpus linguistic methods.

## Participants:

- FAU - Law School (Prof. Dr. Adrian / Michael Keuchen) and FAU - Chair of Computational Corpus Linguistics (Prof. Dr. Evert / Philipp Heinrich / Natalie Dykes) as well as 4 - 8 student assistants.
  - Interdisciplinary project: legal theory and methodology, linguistics, mathematics, computer science
- Funded by the Bavarian State Ministry of Justice

# Research subject of the project

## Court decisions

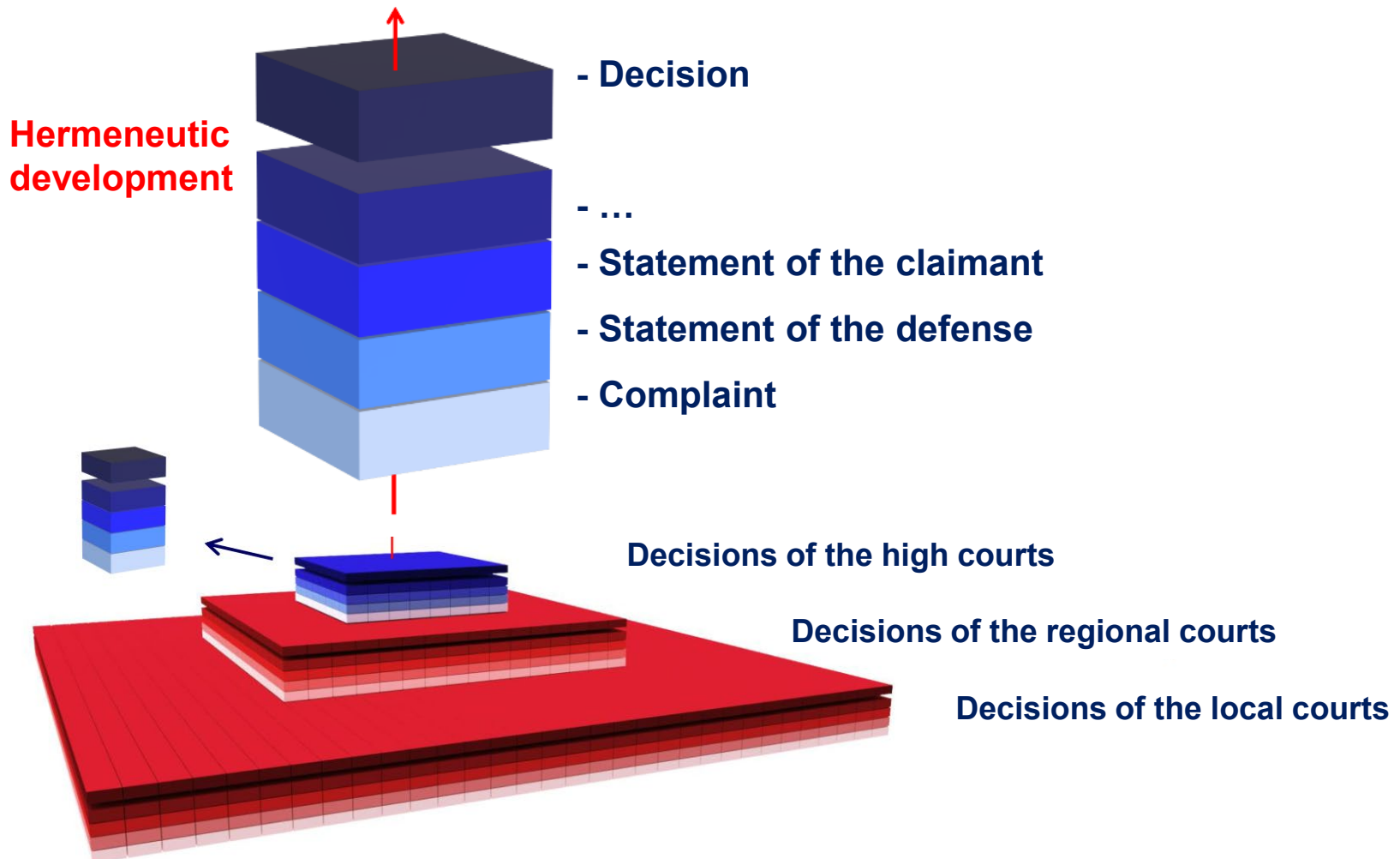
- 604 machine-readable decisions from a bavarian local court concerning traffic accident law and residential tenancy law.
- 1,3 million tokens (words and punctuation) approx. 2,200 tokens per decision.

## Files containing the corresponding legal documents

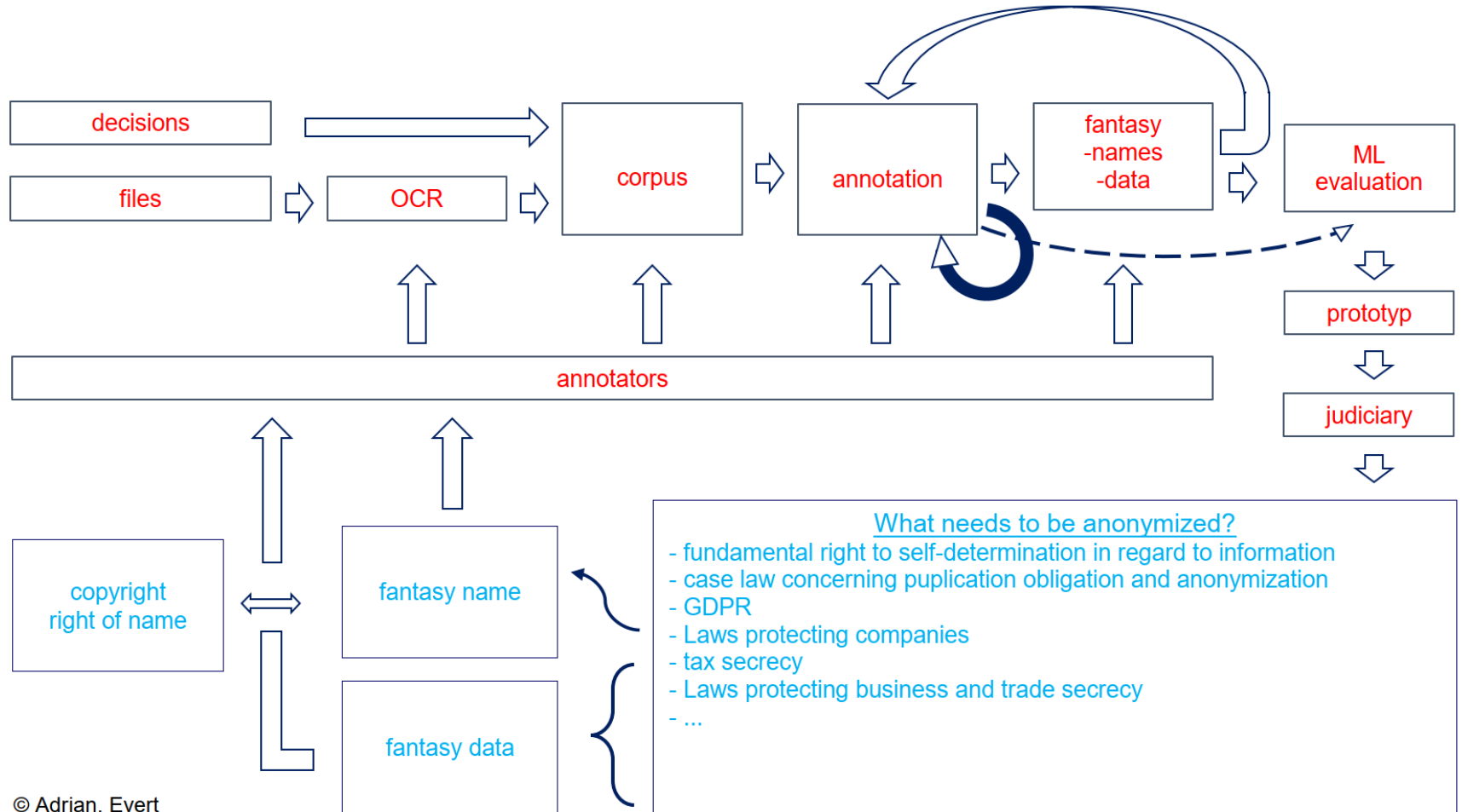
- 557 files (hard copy) are available.
- Scanning only the complaint, statement of defense and other written pleadings regarding the facts of the case.
- Use of text recognition software to achieve machine readability.



# Why files are required?



# Project workflow



## Example: annotation and tagging

The screenshot shows a legal text document with a 'New Annotation' dialog box open. The text in the background is as follows:

38 Ort 90762 Fürth , Gz.: 557/01  
40 wegen Räumung  
42 Ort erlässt das Amtsgericht Hof durch Richter am Sonstiges  
44 ENDURTEIL Name  
46 I. Die Beklagten werden verurteilt, das Anwes  
48 II. Die Kosten des Rechtsstreits werden gegen  
50 III. Das Urteil ist vorläufig vollstreckbar. Die B  
52 Tatbestand:  
54 Die Kläger begehren von den Beklagten die Räumung  
56 Kennzeichen Die Kläger sind Eigentümer des Anwesens Dö  
Ort Mietvertrag (BI. 6/16 d. A.) enthält u. a. folge  
58 Inhalt:  
60 „...  
62 § 2 Nr. 2 Vertragsdauer

The 'New Annotation' dialog box contains the following fields:

- Text:** Beklagten
- Search:** google, wikipedia
- Entity type:** Name (selected), Telefonnummer, Titel, Adresse, Ort, Datum, Kennzeichen, ...
- Entity attributes:**  unsicher, Risiko: ?
- Notes:** (empty)

Buttons for 'OK' and 'Cancel' are located at the bottom right of the dialog box.

# Development of algorithms (logics)

- Annotators produce a gold standard that is used to develop, train and evaluate machine learning algorithms.
- The machine uses this training data to search for patterns in the language in order to identify the text passages which need to be anonymized.
- Based on these algorithms, automatic anonymization may be possible.

The screenshot shows the CQPweb interface for searching Bavarian Court Decisions. On the left is a navigation menu with sections like 'Corpus queries' and 'Saved query data'. The main area is titled 'Standard Query' and contains a search query: `[Herr, Frau, Dr?, Prof?, Zeuge, Richter*] {????*/N}`. Below the query field are settings for 'Query mode' (Simple query (case-sensitive)), 'Number of hits per page' (50), and 'Match strategy' (Standard). A box labeled 'logic' with a green arrow points to the search query field. Below the interface, a snippet of a court decision is shown, with the word 'Zeuge' highlighted in blue. The snippet text includes: '6000 Euro übersteigender Abfindungsbetrag mit an Sicherheit grenzender Wahrscheinlichkeit nicht zusätzlich zur Verfügung stehen würde. Auch hier haben sich die Klägerin und der Beklagte konträr eingelassen und auch der **Zeuge** Adrian hat die Angaben der Klägerin bestätigt. Auch hier kann ein Beweis mit dem Maßstab des § 286 ZPO nicht geführt werden. Insoweit wird auf die Bewertung der Einlassungen

## Outlook and expected outcomes of the project

- Legal foundations for the anonymization.
- Gold standard for the annotation and anonymization of court decisions.
- Possible software prototype for an automatic anonymization.
- Hermeneutic approach with the help of the legal documents.
- Fully anonymous dataset from court decisions and legal documents (after approval by the Bavarian State Ministry of Justice).



# Thanks for your interest!

**Michael Keuchen, [michael.keuchen@fau.de](mailto:michael.keuchen@fau.de)**  
Scientific assistant at the Law School of the FAU



FRIEDRICH-ALEXANDER  
UNIVERSITÄT  
ERLANGEN-NÜRNBERG

FACHBEREICH  
RECHTSWISSENSCHAFT



LEHRSTUHL FÜR  
KORPUS- UND  
COMPUTERLINGUISTIK