DIE INDIVIDUELLE JURISTISCHE SUCHMASCHINE CORE INTERNATIONAL CRIMES DATABASE

Ralph Hecksteden / Jörg Reichert

Geschäftsführer, jurmatix legal intelligence UG (haftungsbeschränkt),
Hauptstraße 28, 66453 Gersheim, DE
li@jurmatix.net; https://legalintelligence.jurmatix.net/

Gesellschafter und Rechtsanwalt, jurmatix Legal Intelligence UG
Hauptstraße 28, 66453 Gersheim, DE
reichert@jurmatix.net; https://legalintelligence.jurmatix.net/

Schlagworte: ***Dokumentmanagement, Wissensmanagement, PDF und Maschinelles Lernen, Individueller Kommentar, Web-Applikation, Alfresco, CMIS, Lucene/Solr***

Abstract: *Die «Core International Crimes Database (CICD)» ist eine webbasierte Dokumentationsplattform und Suchmaschine. Sie bietet einen Zugang zur aktuellen Rechtsprechung zum Völkerstrafrecht und erlaubt den Benutzern, maßgeschneiderte Zusammenfassungen von Rechtstexten zu erstellen. Dazu lassen sich juristische Dokumente annotieren, mit Metadaten versehen und in eine Systematik mit mehr als 8.000 Tatbestandsmerkmalen und juristischen Klassifikationen des Völkerstrafrechts einordnen. Mit den bereits erfassten Annotationen wird ein Machine-Learning Verfahren evaluiert.*

1. Ziel

Die Core International Crimes Database (CICD) ist eine individualisierte Suchmaschine, die für zwei europäische Nichtregierungsorganisationen entwickelt wurde.[1] Sie bietet einen umfassenden Zugang zur historischen und aktuellen Rechtsprechung in den Bereichen Völkerstrafrecht und Menschenrechtsverletzungen seit 1920.

Darüber hinaus erleichtern weiterführende, speziell entwickelte Funktionalitäten dem Benutzer die Erstellung maßgeschneiderter Zusammenfassungen von Rechtstexten. In solchen «Digests» werden Passagen bisheriger Rechtsprechung zu diesem Sachgebiet strukturiert zusammengefasst. Je nach Granulierung der Sachgebietsgliederung ist dabei eine Zuordnung der Rechtsprechung zu Tatbestandsmerkmalen bis auf die Ebene einzelner Beweismittel möglich.

Digests sind ein wichtiges Instrument im Völkerstrafrecht, das auf dem anglo-amerikanischen System des Fallrechts aufbaut. In der Rechtspraxis wird der Digest verwendet, um Fälle mit ähnlichen Fakten oder ähnlicher Problematik aufzufinden um diese dann im Verfahren – wie im Fallrecht üblich – als primäre Rechtsquelle einzuführen. Die reproduzier- und dokumentierbare Erstellung umfassender und gleichzeitig übersichtlicher Digests ist ebenso anspruchsvoll wie arbeitsintensiv. Im Völkerstrafrecht gilt dies, in Anbetracht der hier häufig sehr umfangreichen Urteile, in besonderem Maße. Entsprechend sinnvoll erscheint eine Nutzung informationstechnischer Werkzeuge wie sie durch die CICD in für die Aufgabenstellung maßgeschneiderter Form zur Verfügung gestellt werden.

Zielgruppe der CICD sind sowohl Ermittler als auch Wissenschaftler im Bereich des Völkerstrafrechts. Dem Ermittler wird bei der Suche nach Fakten und Beweisen eine Handhabe gereicht, mit der er gezielt bereits

[1] Dies sind die in Warschau ansässige NGO «Central and Eastern European Initiative for International Criminal Law and Human Rights» (ICLHR, http://www.iclhr.org/en/ [alle Websiten zuletzt besucht im Januar 2018]) und die NGO «Centre for International Law Research and Policy» (CILRAP, https://www.cilrap.org/) mit Sitz in Brüssel.

anerkannten Beweismitteln für Tatbestandsmerkmale nachgehen kann. Zusätzlich kann der Wissenschaftler mit dem Digest retrospektiv die Kohärenz der Rechtsprechung beurteilen.

2. Umsetzung und Datenbasis

Um die manuelle Erstellung der Digests mit den Möglichkeiten moderner Datenverarbeitung zu unterstützen, wurde die CICD von den Verfassern als Mischform aus Autorenwerkzeug und Suchmaschine entwickelt. Dabei werden anwendungsspezifische und teilweise proprietär entwickelte Funktionalitäten wie Annotation auf Satzebene und automatische Übernahme von Textpassagen aus der Legal Tools Datenbank mit bekannten Text- und Suchfunktionen in einer Webanwendung integriert.

Dieses Frontend greift auf die Legal Tools Datenbank[2] des Internationalen Strafgerichtshofs (IStGH) in Den Haag zu, in der im Auftrag des IStGH alle Entscheidungen des Völkerstrafrechts öffentlich zugänglich gemacht werden. Die Entscheidungssammlung beginnt mit den Dokumenten der United Nations War Crimes Commission (UNWCC), die die Gräueltaten des zweiten Weltkriegs dokumentierte und endet bei den jeweils aktuellen Entscheidungen des IStGH. Insgesamt besteht aktuell ein Zugriff auf mehr als 130.000 Dokumente dieses Rechtsgebiets.

Mit dieser Kombination aus umfassender, fachgebietsspezifischer Datenbasis und maßgeschneidertem Such- und Autorenwerkzeug steht eine Komplettlösung für die Arbeit auf dem Gebiet des internationalen Strafrechts zur Verfügung.

3. Rollen Bearbeiter/Leser

Die CICD verfügt vereinfacht über zwei Benutzerrollen: den Bearbeiter und den lesenden Benutzer.

Der Bearbeiter hat Zugriff auf ein Web-Frontend, aus dem er direkt auf die Legal Tools Datenbank zugreifen und Dokumente aus dieser Sammlung in seinen Arbeitsbereich übernehmen kann. Dort kann er dann die Entscheidungen annotieren und mit Metadaten versehen.

Die Metadaten werden vom Bearbeiter dabei auf zwei Ebenen erfasst. Zum einen auf der Dokumentenebene – hierbei werden die Daten meist schon beim Import aus der Legal Tools Datenbank übernommen – und auf Absatz- oder Satzebene. Auf dieser zweiten Ebene werden dann weitere Informationen erfasst, wie:

- Art der Auslegung
- Typ des Arguments (belastend, entlastend …)
- Quelle des Arguments (Zeuge, Gutachten, Bericht …)
- Relation zu anderen Absätzen und Bewertung (unterstützend, widersprechend …)
- Einordnung in die Systematik (siehe 4.)
- Schlagworte

Der lesende Benutzer kann aus den erfassten Daten seinen individuellen Digest erstellen. Dazu steht ihm eine Suchmaschine zur Verfügung, die ihm neben der Volltextsuche über die dokumentieren Absätze auch Filterfunktionen nach Metadaten und der systematischen Einordnung der Absätze gibt.

4. Systematik

Für die Annotation der Entscheidungen steht dem Bearbeiter eine Systematik der materiellen Normen des Römischen Statuts zur Verfügung. In dieser Systematik sind auf fünf Ebenen alle Tatbestandsmerkmale der völkerstrafrechtlichen Tatbestände aufgeschlüsselt. Dies sind im Einzelnen die Tatbestände der Kategorien Genozid, Verbrechen gegen die Menschlichkeit und Kriegsverbrechen (Art. 6 bis 8 des Römischen Statuts).

[2] https://www.legal-tools.org/.

Ebenso stehen die zwölf verschiedenen Modi der persönlichen Verantwortlichkeit in ihren Merkmalen ausgeschlüsselt zur Verfügung. Insgesamt verfügt die Systematik über 8200 Einträge.

Zum Bearbeiten und Erweitern der Systematik wurde in die CICD ein entsprechendes Werkzeug in Form eines Baum-Editors eingebunden. Dieser verfügt über eine Suchfunktion über alle Äste des Baumes, eine Versionsverwaltung und eine Erkennung von Duplikaten.

5. Technische Basis

Die Umsetzung der CICD erfolgte zunächst als Prototyp und unter limitiertem (Zeit-)Budget. Von Beginn an wurde daher angestrebt, möglichst viele frei verwendbare Bausteine einzusetzen, um die Entwicklungskosten gering zu halten und dennoch ein umfangreiches und komfortables Werkzeug erstellen zu können.

5.1. Benutzerschnittstellen

Die CICD-Benutzerschnittstellen sind komplett webbasiert, wodurch beim Redakteur sowie beim Anwender keinerlei Installationsaufwände entstehen. Es handelt sich dabei um «Rich-Clients» auf Basis von HTML 5, die auf verschiedenen JavaScript-Frameworks aufsetzen. Zu nennen sind hier insbesondere:

- jQuery[3] und verschiedene jQuery-Module, die bspw. zur Formularerstellung und Kommunikation mit dem Dokumentmanagement-System im Hintergrund eingesetzt werden,
- PDF.js,[4] mit dessen Hilfe PDF-Dateien innerhalb einer Web-Applikation ohne Rückgriff auf Browser-Plugins dargestellt werden können. Die Anzeige erfolgt ausschließlich über JavaScript und HTML 5. Dadurch wird die Interaktion zwischen Web-Anwendung und PDF-Darstellung ermöglicht – in der Anwendung war diese bspw. zur Markierung und Extraktion von Textstellen erforderlich.

Die Benutzerschnittstellen können in allen gängigen aktuellen Webbrowsern verwendet werden. Die zunehmende Standardisierung und Harmonisierung der HTML 5-Funktionalität macht sich bei der Cross-Browser-Entwicklung dabei positiv bemerkbar.

Auch Performance-Hürden seitens der Web-Browser waren im Zuge der Projektentwicklung nicht zu beobachten, obwohl die dargestellten und bearbeiteten PDF-Dateien teilweise sehr umfangreich sind.

Da die Datenhaltung allerdings serverseitig stattfindet (siehe dazu den folgenden Abschnitt), ist eine gute Netzanbindung zumindest für die Nutzung der Redaktionsoberfläche erforderlich.

5.2. Dokumentmanagement-System

Die Auswahl einer geeigneten Basis zur Datenhaltung war ein längerer Prozess, da verschiedene Open- und Closed-Source-Systeme zum Dokumentmanagement evaluiert werden mussten. Die Anforderungen hieran waren recht hoch, denn folgende Eigenschaften sollte die Plattform bereits bieten, um Entwicklungszeit einzusparen:

- die Abbildung des o. g. Rollenkonzepts (siehe Abschnitt 3),
- die Ablage von umfangreichen PDF-Dokumenten – den kommentierten/annotierten Entscheidungen – nebst frei definierbaren Metadaten,
- die Ablage von Auszügen aus diesen Dokumenten nebst frei definierbaren Metadaten und eine Verlinkung der Auszüge mit ihren Ursprungsdokumenten,
- die Sofort-Indexierung aller hochgeladenen Inhalte mit einer möglichst guten Suchmaschine,
- die Möglichkeit zum Außenzugriff über Webschnittstellen zwecks Anbindung der Oberfläche an das Dokumentenverwaltungssystem,
- eine Versionsverwaltung zwecks Nachvollziehbarkeit von Änderungen an Bearbeitungen.

[3] https://jquery.com/.

[4] https://mozilla.github.io/pdf.js/.

Nach Evaluation einiger Systeme fiel die Wahl auf «Alfresco»[5], ein weit verbreitetes Dokumentmanagement-System, das in einer kostenlosen Community Edition angeboten wird.

Die Software erfüllt die oben genannten Eigenschaften vergleichsweise gut. Jedoch ist die für die Community-Edition verfügbare Dokumentation recht dürftig, was die Nutzung von Schnittstellen sowie die Entwicklung von Zusatzfunktionalität erschwert.

Die grundsätzlich vorhandene Möglichkeit, Inhalten selbst definierbare Metadaten beizustellen und diese auch zu durchsuchen und zu filtern, ist dadurch erschwert, dass nachträgliche Änderungen am Metadatenmodell – bis auf Hinzufügungen – nicht möglich sind.

Insgesamt ist Alfresco aus Sicht der Verfasser jedoch ein Dokumentmanagement-System, das zum Aufbau juristischer Knowledge-Management-Lösungen gut einsetzbar ist. Positiv hervorzuheben ist dabei, dass die Software den CMIS-Standard[6] zum Datenaustausch zwischen/mit Dokumentmanagementsystemen unterstützt, der auch im Rahmen der hier vorgestellten Lösung verwendet wird.

Für das System spricht auch, dass es seinerseits auf bewährten Open-Source-Modulen, wie dem Webapplikations-Server Apache Tomcat,[7] der Datenbank PostgreSQL[8] oder der Suchmaschine Lucene/Solr[9] aufbaut. Letzteres war bei Umsetzung der CICD-Anwendung hilfreich, denn für einige Suchanfragen war der direkte Durchgriff auf Lucene/Solr erforderlich und möglich.

5.3. Öffentliches Webinterface

Die öffentliche Abfrageoberfläche für den nur-lesenden Benutzer wurde aus Sicherheitsgründen gekapselt umgesetzt, so dass keine direkte Kommunikation zwischen dem Webbrowser und dem Dokumentmanagement-System im Hintergrund stattfindet. Andernfalls hätten Benutzername und Passwort in der Webapplikation hinterlegt werden oder ein anonymer Gastzugriff auf das Dokumentmanagement-System erlaubt werden müssen.

Der Webserver des Dokumentmanagement-Systems wurde dazu um eine kleine Java-Webapplikation ergänzt, die intern über die vorhandenen Alfresco-Schnittstellen sowie direkt mit der Suchmaschine Lucene/Solr kommuniziert und nach außen einfache Webservices für Suchen und Dokumentabrufe bereitstellt. Es werden dabei nur Annotationen zu Dokumenten durchsucht, die von Redakteuren freigegeben wurden.

6. Automatisierung mittels Machine-Learning

In der CICD wurden von den Anwendern bereits mehrere tausend relevante Textstellen annotiert. In Anbetracht dieser großen Anzahl kategorisierter und mit Metadaten versehener Auszüge liegt die Möglichkeit nahe, damit maschinelle Verfahren zu trainieren, um dann neu erfasste Dokumente textlich zu analysieren und den Bearbeitern potenziell zu annotierende Abschnitte der Dokumente vorzuschlagen. Hierzu werden derzeit verschiedene Machine-Learning-Verfahren evaluiert.

6.1. Aufbereitung der PDF-Datenbasis

Die zu annotierenden Entscheidungen des Völkerstrafrechts liegen im PDF-Format vor. Das für den menschlichen Benutzer meist angenehm zu lesende Format stellt für die maschinelle Weiterverarbeitung oft große Hürden dar. PDF-Dateien sind – bis auf neuere barrierefreie PDFs – layoutbasiert und bieten nahezu keine semantische Information. Ob ein Textbereich bspw. eine Überschrift, einen Absatz oder eine Fußnote darstellt, ist bei der Text-Extraktion nicht ohne weiteres automatisiert zu erkennen.

5 https://www.alfresco.com/de/, verwendet wird die Version 5.1.
6 https://chemistry.apache.org/project/cmis.html.
7 http://tomcat.apache.org/.
8 https://www.postgresql.org/.
9 https://lucene.apache.org/solr/.

Darüber hinaus entspricht die Reihenfolge der extrahierten Zeichen und Wörter je nach verwendeter Software zur Erzeugung der PDF-Dateien nicht immer der Lese-Reihenfolge oder Wörter sind durch Leerzeichen oder Zeilenumbrüche zerrissen (siehe Abbildung 1).

finding or where the finding is wholly erroneous.[66] The Appeals Chamber has adopted the statement of general principle contained in the ICTY Appeals Chamber decision in *Kupreškić et al.*, as follows:

s:
finding or where the findi
ng is wholly erroneous.
66
The Appeals Chambe
r has adopted the statement of general principle contained in the ICTY Appeals Chamber decision in Kupreški
c et al., as follow

Abbildung 1: PDF-Absatz und fehlerhaft extrahierter Text

Das stellte bereits für die Text-Extraktion der markierten und annotierten Passagen eine recht große Herausforderung dar. Und auch das Auslesen der Absätze neu hinzugefügter Dokumente, die von maschinellen Verfahren analysiert werden sollen, ist hierdurch erschwert. Verfügbare Programme zum Auslesen von Texten aus PDF-Dateien konnten daher nicht ohne Anpassung eingesetzt werden und mussten um Verfahren zum Löschen von Fußnotenbereichen und Zusammensetzen von Absätzen, die durch Seitenumbrüche zerrissen werden, ergänzt werden.

6.2. Trainings-Sets und testweise Anwendung auf Vergleichsdaten

Zur Prüfung der Tauglichkeit maschineller Verfahren zur Unterstützung der Dokumentation wurde ein Trainings-Set aus ca. 2'000 der bereits markierten relevanten Textstellen erzeugt. Es wurden Annotationen ausgewählt, die von den Dokumentaren mit der Kategorie «Element of Crime» versehen wurden, was einen Bezug zu einem Tatbestandsmerkmal bedeutet.

Die Auszüge einiger bereits annotierter Dokumente wurden dabei ausgespart, um diese Dokumente als Vergleichs-Set für die Vorhersage des automatischen Verfahrens zur Einordnung von Absätzen verwenden zu können.

Als Verfahren werden derzeit ein Naive-Bayes-Klassifikator[10] und ein Normalized-Compression-Distance-Algorithmus[11] getestet.

Zum Zeitpunkt der Einreichung dieses Beitrags wurden mit den evaluierten Verfahren noch keine Ergebnisse erzielt, die eine ausreichend hohe Vorhersagequalität für eine automatisierte Dokumentations-Unterstützung erreichen. Textstellen aus der Vergleichsmenge, die Dokumentare als relevant für die o.g. Kategorie «Element of Crime» eingestuft haben, wurden von den Algorithmen nicht ausreichend oft als «sehr ähnlich» zu der «erlernten» Datenbasis eingestuft.

10 https://de.wikipedia.org/wiki/Bayes-Klassifikator.
11 https://en.wikipedia.org/wiki/Normalized_compression_distance.

Allerdings wurden die Absätze zumindest im Durchschnitt deutlich ähnlicher zum Trainigs-Set eingestuft als die von den Dokumentaren nicht markierten Textstellen der Vergleichsdokumente (siehe Abbildung 2).

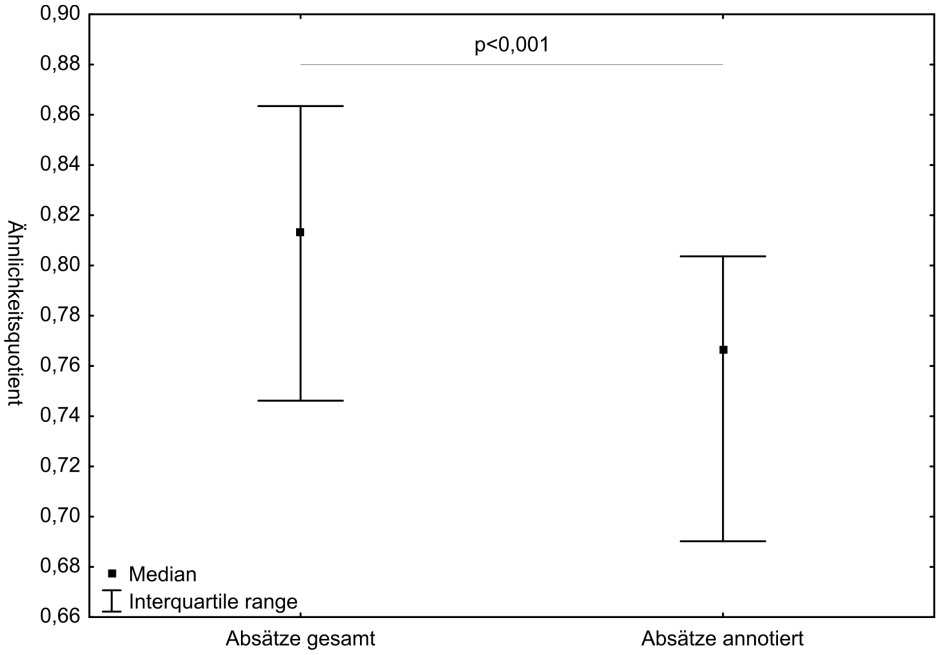

Abbildung 2: Verteilung der Ähnlichkeitswerte eines Vergleichsdokuments

Die evaluierten Verfahren würden sich daher zumindest für andere Hilfestellungen eignen. Zu denken ist hier insbesondere an eine Berücksichtigung der Ähnlichkeitswerte bei der Aufbereitung der Dokumente für die Suche und Relevanzsortierung.

7. Fazit

Durch Verwendung frei verfügbarer Module war es recht schnell möglich, ein komfortables webbasiertes Knowledge-Management-System zur Erschließung der Rechtsprechung zum Völkerstrafrecht zu implementieren. Das PDF-Format, das im juristischen Umfeld oft anzutreffen ist, stellt aber große Hürden bei der Extraktion und Weiterverarbeitung von Informationen dar. Verfahren zur Kategorisierung von Texten, die auf maschinellen Verfahren basieren, bringen «out of the box» zwar interessante aber nur sehr bedingt verwertbare Ergebnisse.