TEXTANALYSE IM KONTEXT DER RUNDFUNKANSTALTEN

FORSCHUNG UND PRAXIS

Dr. Jens Fisseler & Dr. Joachim Köhler | 10.10.2019 | NETZWERK MASCHINELLE VERFAHREN IN DER ERSCHLIESSUNG | Frankfurt

Fraunhofer IAIS

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Intelligent Systems that Work!

Über 300 Wissenschaftler*innen

Über 180 Forschungs- & Industrieprojekte pro Jahr

Mehrere Jahrzehnte Background in KI

Exzellenzforschung im Bereich Künstliche Intelligenz, Machine Learning & Big Data

Bereitstellen von Technologien & Services zum Aufbau intelligenter Systeme

Ausbildung der nächsten Generationen von KI- und Data Scientists

Beratung und Unterstützung von Unternehmen und öff. Organisationen bei der digitalen Transformation

Fraunhofer IAIS

Fraunhofer IAIS
Erfahrungen, Kompetenzen, Technologien im Medienumfeld

- Langjährige Erfahrung (seit 2002) zur Forschung & Entwicklung von Medientechnologien, Schwerpunkt Bild/Video/Audio/Dokumenten-Analyse (KI, ML)
  - Entwicklung einer führenden Audio Mining Lösung
  - Sprachassistenten
  - Bild- und Videoerkennung
  - Text-/Dokumentenanalyse: OCR/Layout-Erkennung
  - Entwicklung einer skalierbaren Mining-Plattform
- Ausgewählte Kunden & Partner: WDR, DW, ProSiebenSat.1, ZDF, ZDF-Digital, ARTE, ARD-Mediathek, NZZ, etc.
- Konzeption und Entwicklung der Deutschen Digitalen Bibliothek (DDB)
- Viele Fachvorträge (IRT, VFM, IBC, IFTA, Global Media Forum, DDB)

Langjährige Erfahrung und tiefe Vernetzung in die deutsche Medienbranche

Content in der ARD/WDR:

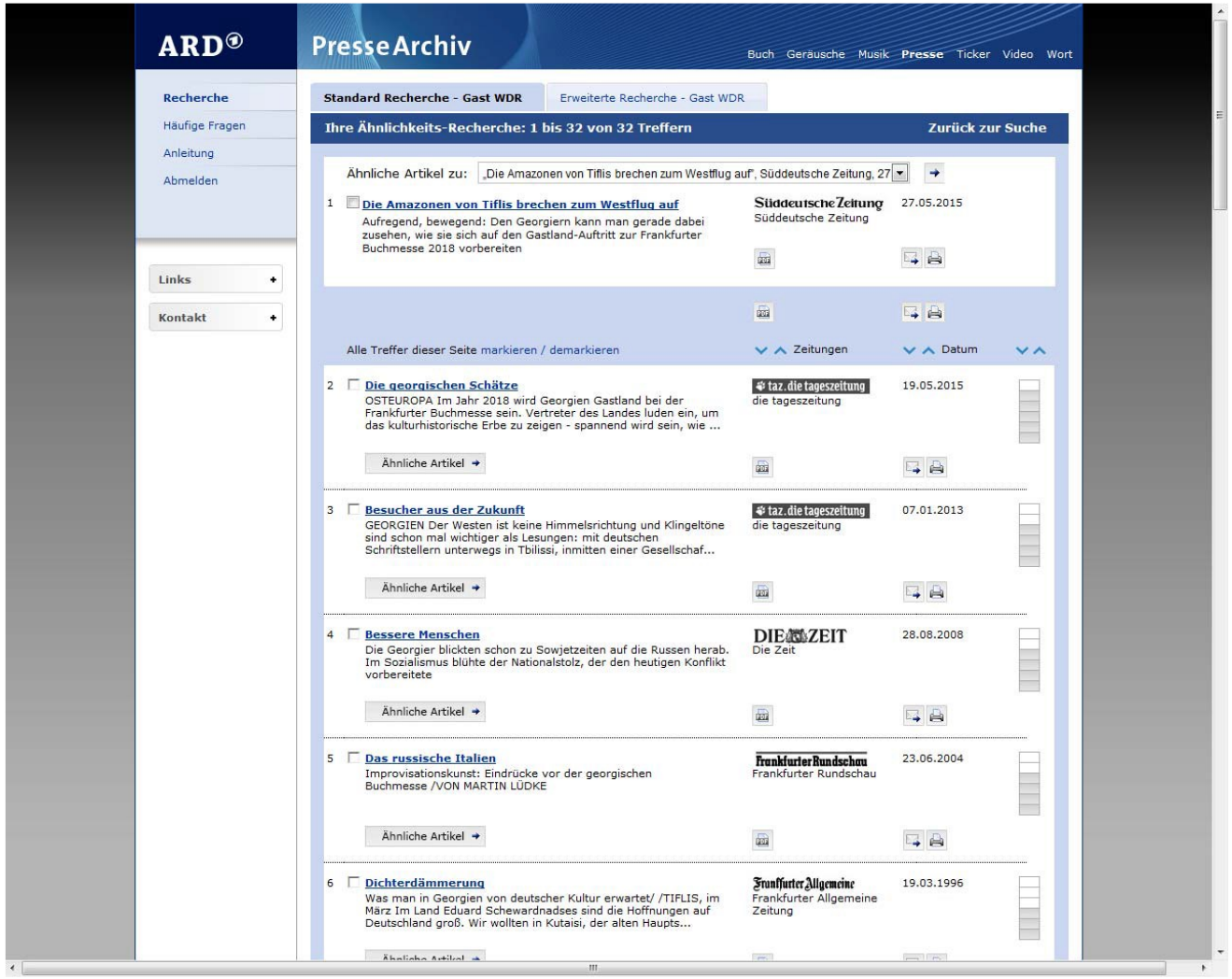
- **Text-Mining Anwendungskontexte**
- **AV-Inhalte mit Speech-to-Text (Audio Mining)**

Fraunhofer IAIS

Text-Mining Anwendungskontexte im Broadcast-Umfeld

PAN Pressedatenbank

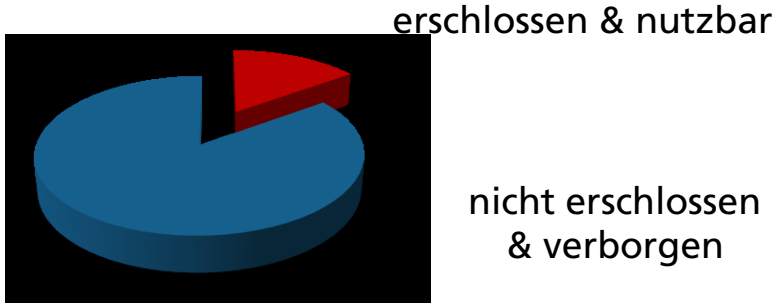
- Pressearchivdatenbanken als Recherchetool
  - Pressearchiv für die Recherche von AV-Produktion
  - ARD-Pressearchiv (PAN) enthält ca. 40 Mio. Artikel aus Zeitungen und Zeitschriften
  - Täglich werden ca. 10.000 Artikel neu in den PAN-Bestand aufgenommen und semi-automatisch mittels NER und ARD-Sachklassifikation verschlagwortet
  - Die ARD-Sachklassifikation enthält ca. 10.000 vorgegebene und ca. 10.000 freie Deskriptoren
  - PAN-Klassen: ca. 2300 Klassen, hierarchisch organisiert



© Westdeutscher Rundfunk / Dokumentation und Archive /

Erschließung von AV-Beiträgen in der ARD

- Erschließung von AV-Inhalten (Radio- und Fernsehbeiträge)
  - Wiederverwendung von bereits produzierten und gesendeten Beiträgen (schnelle und genaue Recherchierbarkeit)
  - Archivsysteme (ARCHIMEDES, FESAD, HFDB)
  - Neues Crossmediales Metadatensystem: medas / MDH
  - Vorverarbeitung (Umwandlung von Audio zu Text) mittels Audio Mining
- Metadaten
  - Keywords und Transkripte
  - Entitäten und semantische Konzepte
  - Personen, Sprecher (ARD Normdatenbank)
  - Zusammenfassungen

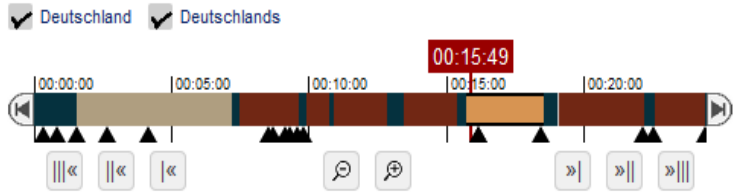


Automatische Spracherkennung

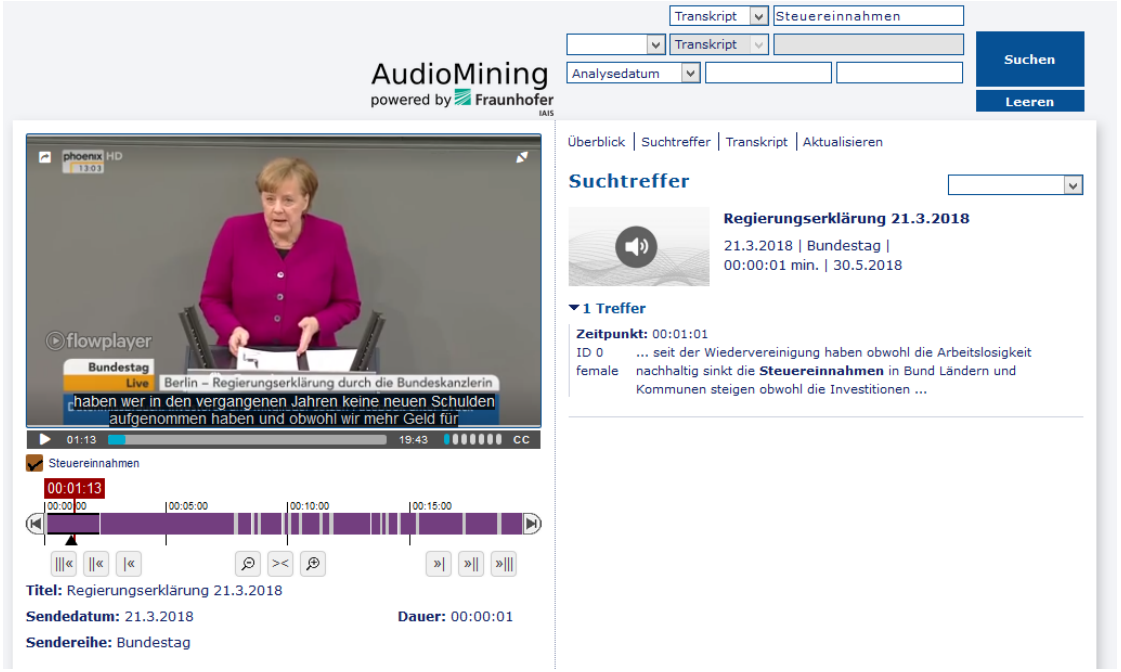
Zeitpunkt: 0:00:12
ID 1 male ... an Bundeskanzlerin Merkel geschrieben die Überschrift **Deutschland** ist ein Überwachungsstaat ...

Zeitpunkt: 0:00:44
ID 1 male ... Sommerpressekonferenz haben sie gesagt **Deutschland** sei kein Überwachungsstaat seit den ...

Strukturierte Aufbereitung



Fraunhofer IAIS Audio Mining



- Ermöglicht Suche anhand der gesprochenen Wörter in Audio- und Videodateien
- Ermöglicht ein schnelles Navigieren durch Audio- und Videodateien mittels Segmentierung

Funktionalität

- Strukturanalyse
  - Audiosegmentierung
  - Sprecherclustering / Sprechererkennung
- Automatische Spracherkennung
- Postprocessing (Punktuation, Zahlen)
- Durchsuchbarkeit der Audio- und Videodateien

Speicherung als Metadaten Indexierung

Text-Mining-Services in der Mining Platform

Orchestrierung der KI-Services in der MDH Mining Platform (medas)

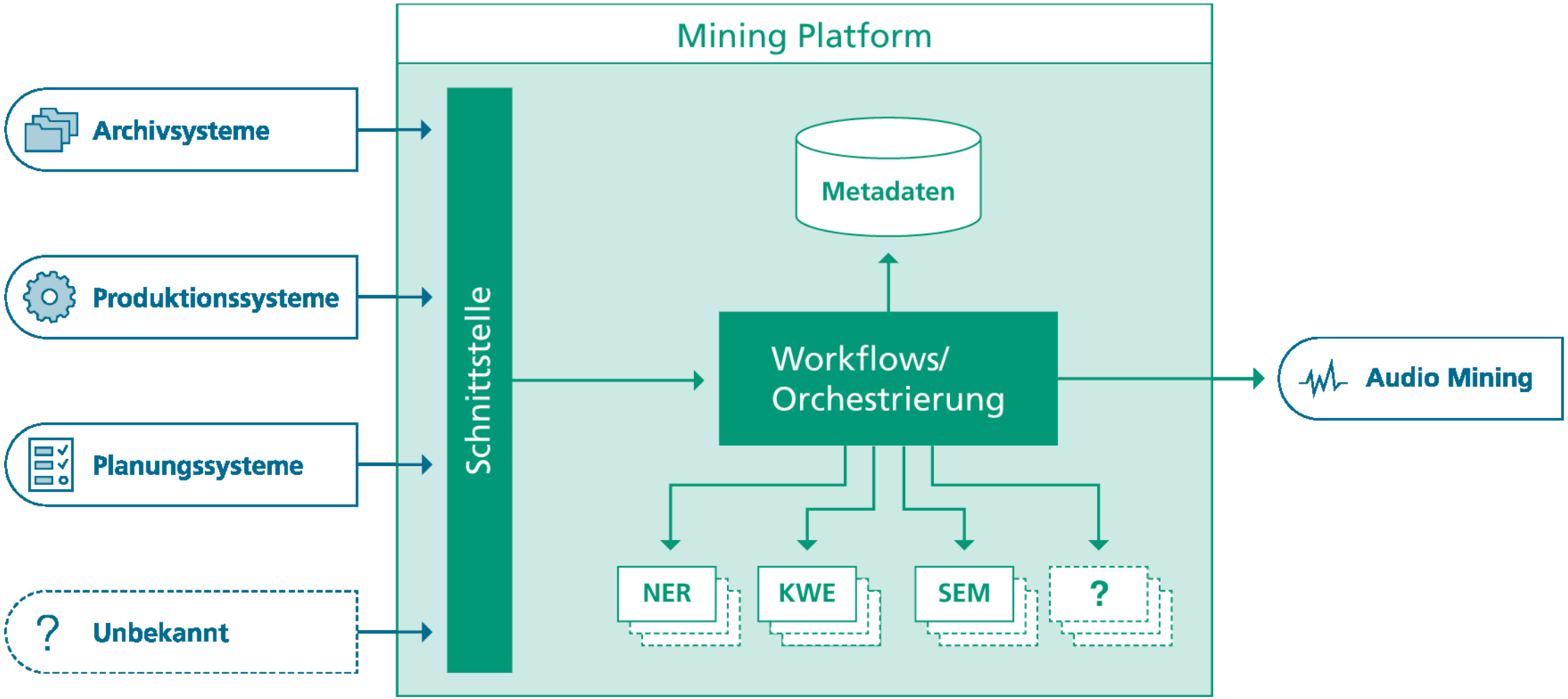
Ziele

- Multimediales Mining-System zur automatischen Metadaten-Generierung
- KI-basierter Metadaten-Mining-Services für Text, Bild, Audio und Video
- Einbindung und Nutzung von ARD-Wissensquellen
- Workflow-Orchestrierung der Mining-Services
- Zentraler Zugang für alle Systeme der ARD und Drittsysteme

Vorteile

- Nutzung des ARD Wissens bei der Modellerstellung und Training führt zu besseren Ergebnissen
- Anpassbar an die Erfordernisse der ARD (Sprache, Wissen)
- Tiefe Vernetzung der erzeugten Metadaten möglich
- Technologie- und Datensouveränität

Schematischer Aufbau der Mining Platform



Nächste Entwicklungsschritte

- Offene Lösung ermöglicht Einbindung beliebiger Mining-Services
- Geplante Integration weiterer Mining-Services
  - Gesichts- und Personenerkennung
  - Topic-Modelling
- Verbesserung bestehender Mining-Services
  - NER-Disambiguierung und -Linking
  - Key-Phrasen
  - Audio-Mining: Sprechererkennung, Spracherkennung, Transkription Englisch

Langfristig: Video-Mining-Services

KI für die Textanalyse

- Keyword Extraction
- Named Entity Recognition (NER)
- Semantic Tagging
- Topic Modelling

Text-Mining-Service „Keyword Extraction“

Extraktion von statistisch signifikanten Wörtern (Wörter, die in einem Text häufiger vorkommen als in einem Gesamtkorpus von Texten)

Die Frau, die wie ein Computer rechnete.

Artikel-Inhalttext: Von Kristina Vaillant

Als Katherine Johnson im Sommer 1953 beim Langley Research Center, der amerikanischen Raumfahrtbehörde in Hampton im Bundesstaat Virginia, anfing, war sie eine von mehreren Hundert meist weiblichen Mathematikerinnen. Computers nannte man sie. In drei Schichten an sechs Tagen in der Woche rund um die Uhr lieferten sie mit ihren mechanischen Rechenmaschinen Daten für die Aerodynamik-Experimente der Ingenieure im Windkanal.

[…]

Schlüsselwörter: Raumfahrt, Mathematik, Raumfahrtbehörde, Astronaut, Rechenassistenten, Luftfahrt, Mathematiktalent, Vektorgeometrie, Orbitalflug, Buch, Rüstungsindustrie, Filmproduktion, Fortran, Naca, Bremsraketen, Afroamerikanerinnen

Quelle: https://vaillant-texte.de/katherine-johnson-aus-dem-film-hidden-figures-zum-100-geburtstag/

Text-Mining-Service „Keyword Extraction“

- Automatische Extraktion von relevanten Schlüsselwörtern aus einem gegebenen Dokument
- Methodik: TF-IDF (Term Frequency–Inverse Document Frequency)
- Grundidee: Gewichtung eines Begriffs anhand der Häufigkeit seines Vorkommens im gegebenen Dokument in Bezug auf eine Gesamtdokumentenmenge (Korpus)
- Mögliche Anwendungen:
  - Als Gewichtung von Dokumenten bei der Volltextsuche
  - Als Suchfilter
  - Als Input für komplexere Verarbeitungsschritte (z. B. Semantic Tagging)

$$w_d(t) = f_{t,d} \times \log\left(\frac{N}{n_t}\right)$$

$w_d(t)$ – Gewicht des Terms t im Dokument d

$f_{t,d}$ – Anzahl der Vorkommen von t in d

N – Gesamtanzahl der Dokumente

n_t – Anzahl der Dokumente, die t enthalten

Text-Mining-Service „Named Entity Recognition (NER)“

Identifikation von Entitäten (derzeit Personen, Organisationen, Geografika)

Die Frau, die wie ein Computer rechnete.

Artikel-Inhalttext: Von [Kristina Vaillant^{PER}]

Als [Katherine Johnson^{PER}] im Sommer 1953 beim [Langley Research Center^{ORG}], der amerikanischen Raumfahrtbehörde in [Hampton^{LOC}] im Bundesstaat [Virginia^{LOC}], anfing, war sie eine von mehreren Hundert meist weiblichen Mathematikerinnen. Computers nannte man sie. In drei Schichten an sechs Tagen in der Woche rund um die Uhr lieferten sie mit ihren mechanischen Rechenmaschinen Daten für die Aerodynamik-Experimente der Ingenieure im Windkanal.

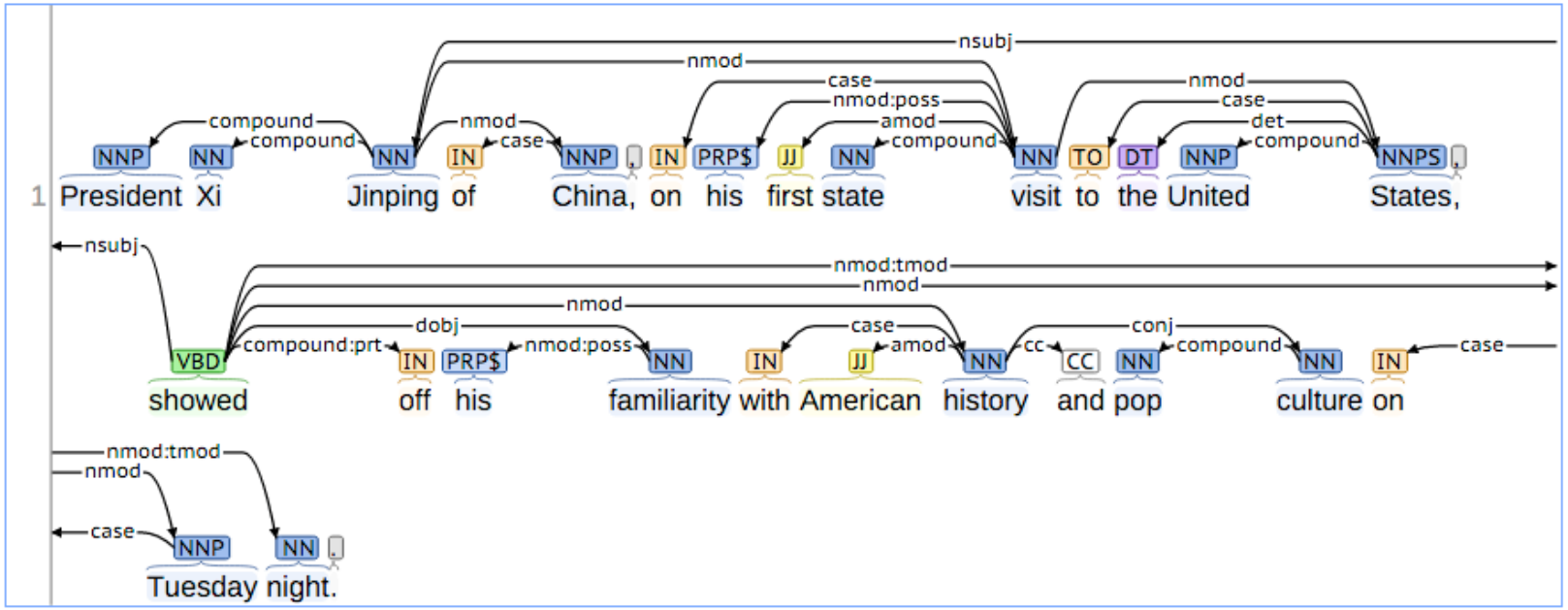
[…]

Für die Ingenieure, die beim [National Advisory Committee for Aeronautics^{ORG}], kurz Naca, der Vorgängerorganisation der [Nasa^{ORG}], die wissenschaftlichen Grundlagen für die Luftfahrt erarbeiteten, waren [Johnson^{PER}] und ihre Kolleginnen namenlose Hilfskräfte. […]

Quelle: https://vaillant-texte.de/katherine-johnson-aus-dem-film-hidden-figures-zum-100-geburtstag/

Text-Mining-Service „Named Entity Recognition (NER)“

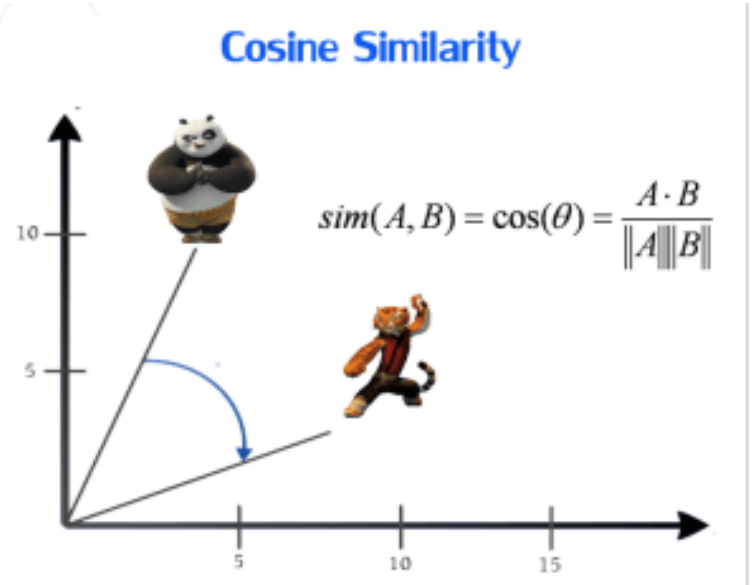
- Automatische Erkennung von Entitäten in einem gegebenen Dokument
- Technologie: Stanford Named Entity Recognizer (integriert in DKPro)
- Technische Grundlagen: Part-of-speech-Tagging (POS Tagging) + Linear Conditional Random Fields
- Anbindung der State-of-the-Art-Implementierung in Java
- Mögliche Anwendungen:
  - Zur Erfassung in Datenbanken
  - Als Kandidaten für Keyphrase Extraction



Bildquelle: https://stanfordnlp.github.io/CoreNLP/

Textanalyse mittels KI

- Verwendung von Deep Learning im Textverstehen
- Word-Embeddings: Umwandlung von Wörtern in Vektoren
- Worte mit ähnlicher Verwendung haben ähnliche Vektoren
- Ähnlichkeit wird durch Cosinus-Abstand berechnet
- Word2Vec und ähnliche Methoden erlauben Abstände und Ähnlichkeiten zwischen Wörten, Phrasen und Dokumenten zu berechnen

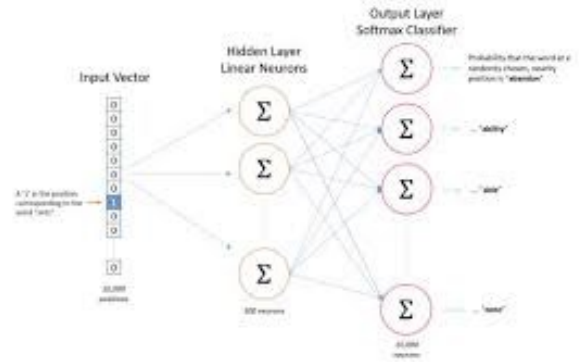


	Trump	earned	an	economics	degree	from	Wharton
100-dim. embedding	0.1,0.4,1.5	1,2,2.0,0.5	0.7,0.1,0.3	2.6,1.2,1.3	5.7,1.4,0.5	2.1,2.8,1.1	0.2,0.3,1.4

Anwendungsfall Semantic Tagging
Kontext PAN Archiv

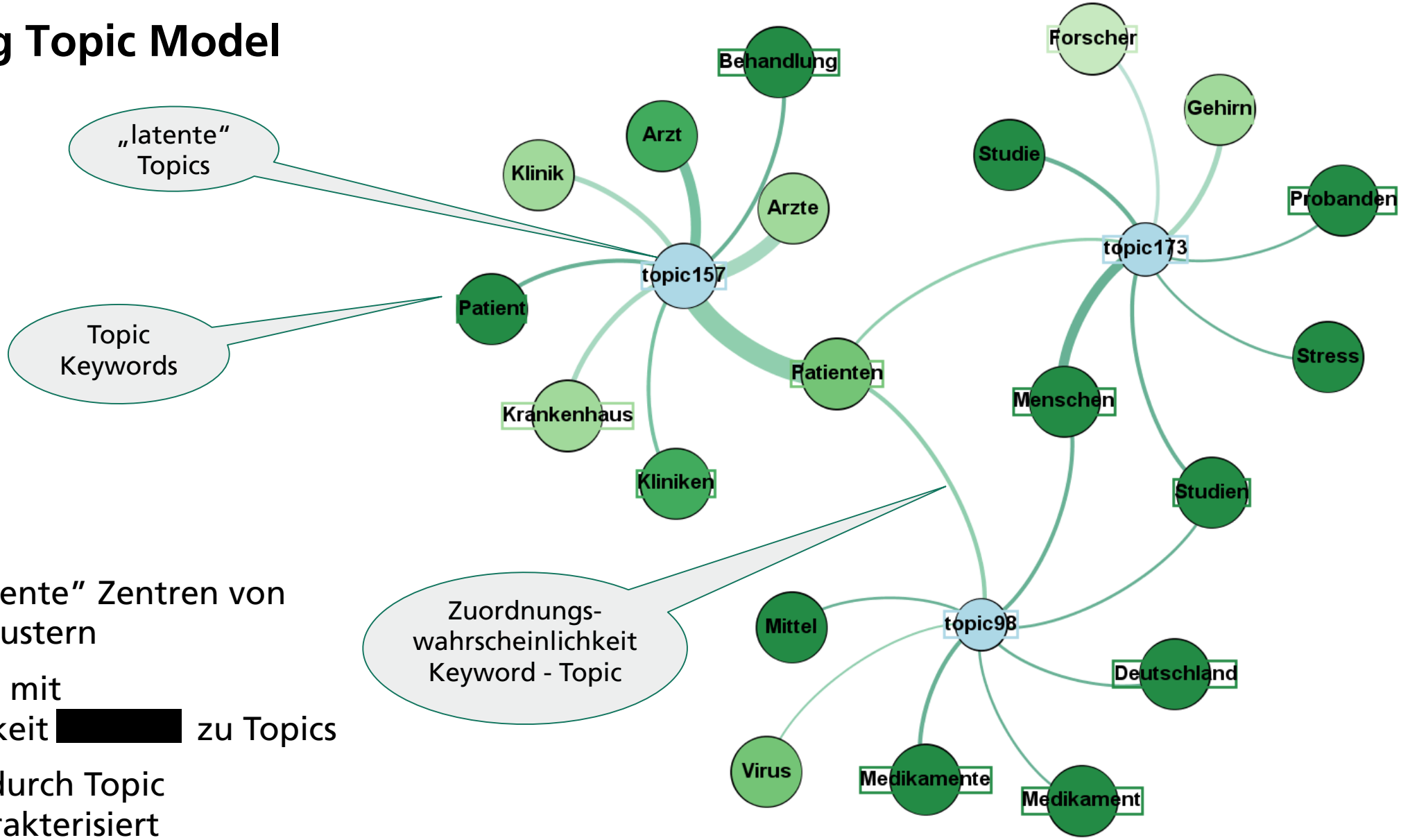
- Nutzung des StarSpace-Modells (Facebook)
- Modifiziertes Word2Vec Verfahren auf Basis von Annotationen
- Trainiert mit Daten und Annotationen des PAN-Pressearchivs
- Modell versieht neue Texte mit passenden Annotationen

ID	PRESSARTICLE_DOCID39868521_CHGID…
CONTENT	Immer noch reisen Tausende Flüchtlinge über Serbien nach …
KEYWORDS	»Refugee_Aid_Serbia« Beruf_Unternehmer…
LOCATION	Kroatien Serbien …
ORGANIZATION	Vereinten_Nationen
PERSON	Kos Nenad_Popovic Popovic Paul_Simon
TOPIC	topic_50
TOPIC KEYWORDS	Flüchtlinge Deutschland …



https://cw.fel.cvut.cz/old/_media/courses/xp36vpd/vpdstarspace.pdf

Visualisierung Topic Model

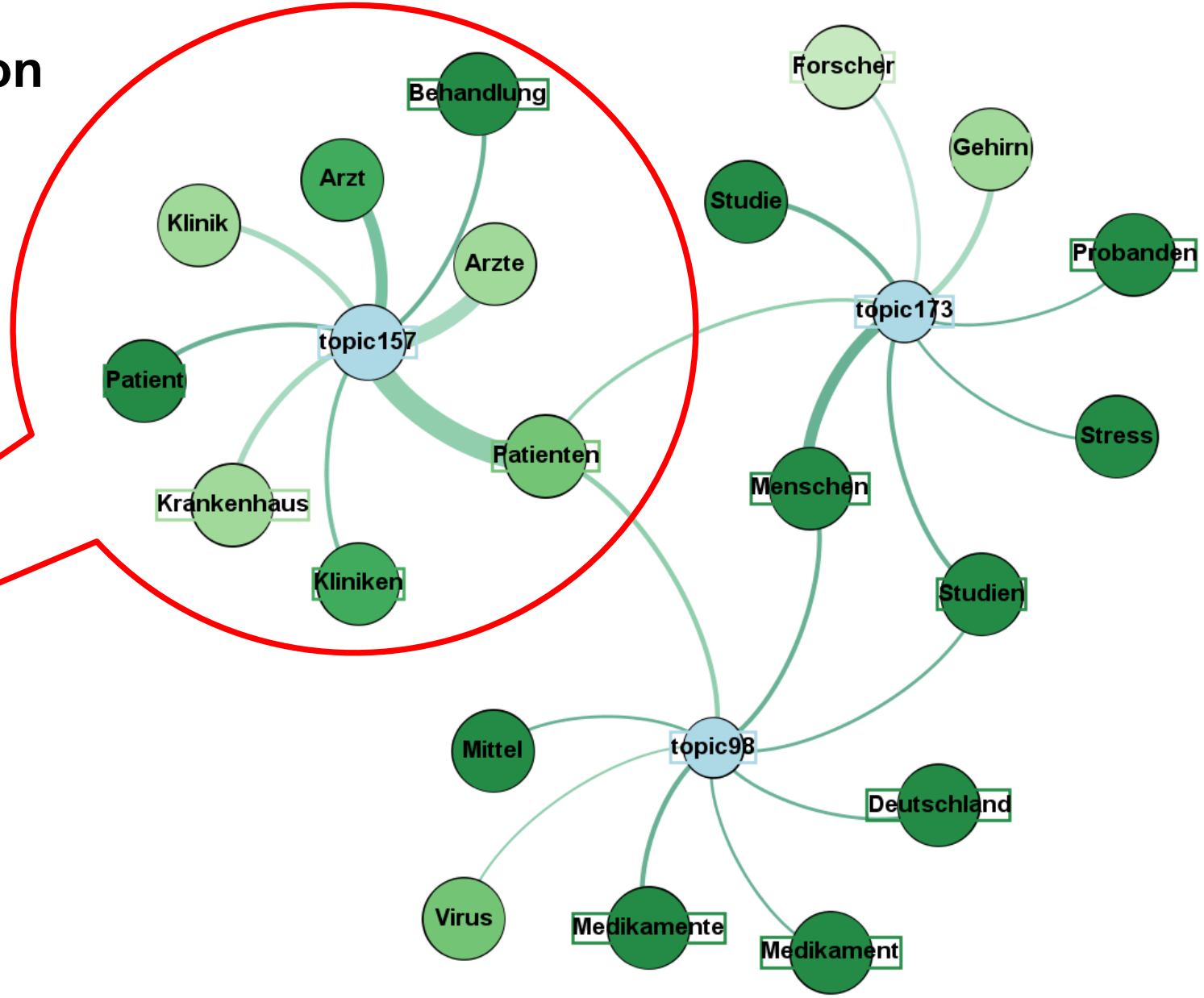


- **Topics** sind “latente” Zentren von Dokumenten-Clustern
- **Worte** gehören mit Wahrscheinlichkeit [redacted] zu Topics
- Topics werden durch Topic **Keywords** charakterisiert

Fraunhofer IAIS

Topic-Nr. & Keyword Annotation

ID	PRESSARTICLE_DOCID39868516_CHGID00000000000029505073201801271016 14
CONTENT	Die Krankenkassen erhöhen ihre Zusatzbeiträge und Zuzahlungen steigen. Was sind die Gründe? Zunächst möchte ich die positiven Gründe nennen: den medizinischen Fortschritt und neue Behandlungsmethoden, die den Menschen helfen. Das kostet Geld. Aber die Politik hat auch teure Reformen verabschiedet, die Krankenkassen und Versicherte stärker belasten. Und wir haben nach wie vor große Steigerungen bei den Arzneimittelausgaben, die auch dazu beitragen, dass die Zusatzbeiträge erhöht werden müssen. Sie schieben den Schwarzen Peter der Politik zu, denn die hat teure Gesetze gemacht, die ja noch nicht alle ihre Wirkung entfaltet haben? Ich finde, man muss Ross und Reiter benennen. Krankenhausreform, Präventionsgesetz, Hospiz- und Palliativgesetz sind zum Teil auch positiv zu bewerten, kosten aber …
TOPIC	topic_157
TOPIC KEYWORDS	Patienten Ärzte Arzt Klinik Krankenhaus Patient Behandlung Kliniken



Fraunhofer IAIS

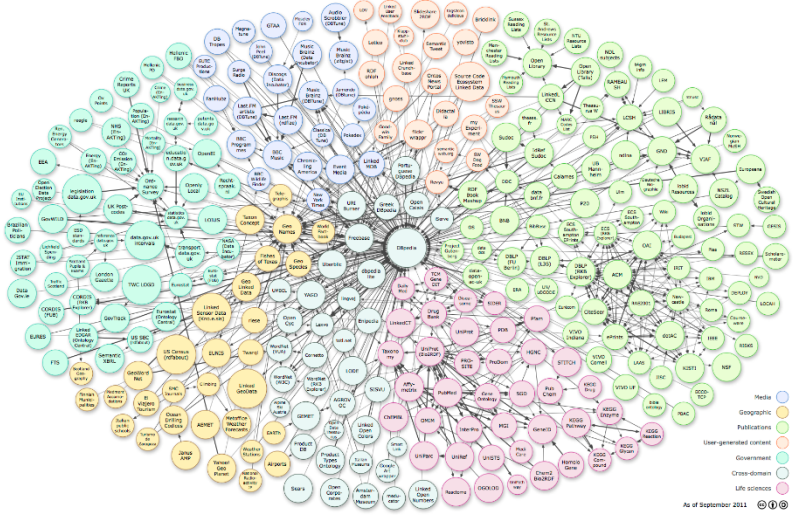
Ausblick

Disambiguierung und Verlinkung von Entitäten

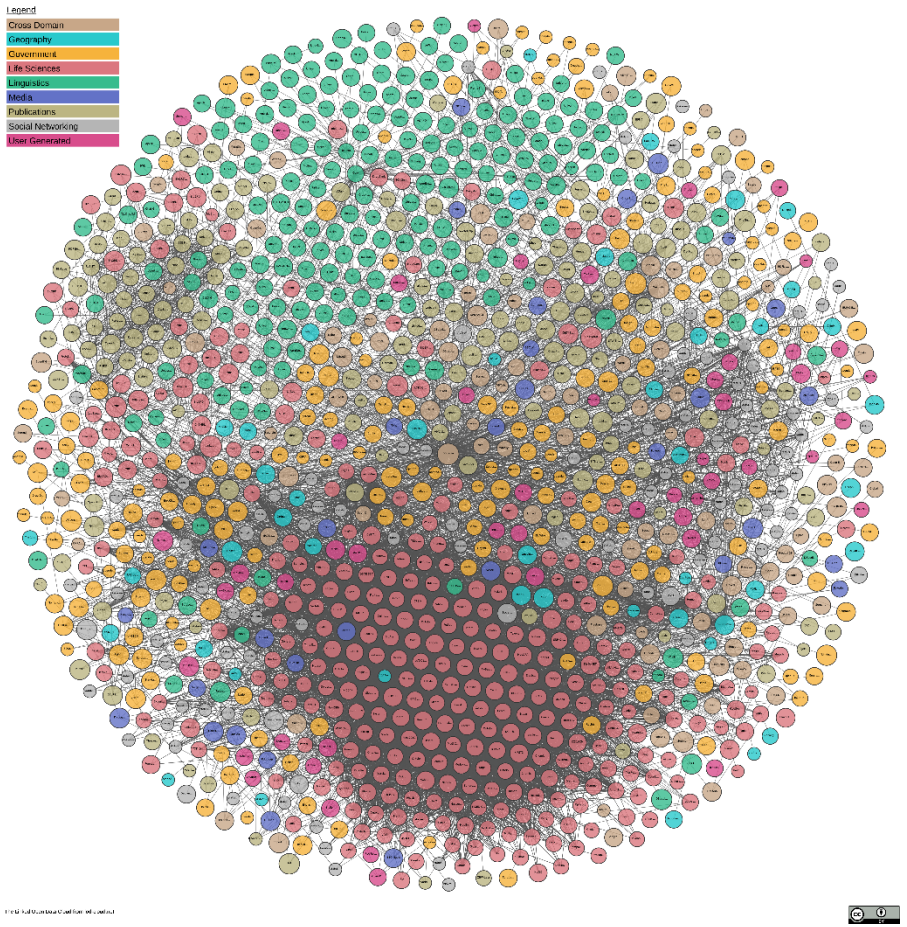
In Planung:

- Named Entity Disambiguation (NED)
- Named Entity Linking (NEL)
- Auf der Basis von öffentlich zugänglichen Quellen, z. B. Linked Open Data Cloud (LOC Cloud, https://lod-cloud.net)

2011



2019



Fraunhofer IAIS

Zusammenfassung und Ausblick
Textanalyse im Kontext der Rundfunkanstalten

- Sehr große Datenmengen erfordern automatisierte Verfahren zur Inhaltserschließung
- Herausforderungen: Manuell versus automatisch erstellte Klassen und Deskriptoren (Qualität, Pflegeaufwand, Kosten, Anwendungskontexte)
- Fraunhofer IAIS verfügt über produktive, skalierbare und anpassbare Module zur
  - Keyword Extraction
  - Named Entity Recognition (mittels DKPro)
  - Semantic Tagging
  - Topic Modelling (mittels DKPro)
- Weitere Innovationen
  - Robuste Disambiguierung, Verwendung von Wissensgraphen
  - Erschließung für audiovisuelle Inhalte

Intelligent Systems that Work!

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Kontakt

Dr. Joachim Köhler

Abteilungsleiter NetMedia

+49 (0)2241 14-1900

joachim.koehler@iais.fraunhofer.de

www.iais.fraunhofer.de

Dr. Jens Fisseler

Fachkoordinator Software & Content Analysis

+49 (0)2241 14-1974

jens.fisseler@iais.fraunhofer.de

Fraunhofer IAIS