Tagungsbericht

Marlies Ockenfeld*

Zukunft für Dokumentare

4. h_da Symposium zur wissenschaftlichen und Mediendokumentation 2019

<https://doi.org/10.1515/iwp-2019-2063>

Bereits zum 4. Mal fand am 6. November 2019 das h_da Symposium zur wissenschaftlichen und Mediendokumentation auf dem Medien-campus Dieburg der Hochschule Darmstadt statt. Die von Professor Geribert E. Jakob organisierte Veranstaltung findet im Rahmen der Weiterbildung zum Wissenschaftlichen Dokumentar (WissDok Volontariat) statt, steht aber auch der Fachöffentlichkeit offen. 23 angehende Forschungsdatenmanager und Mediendokumentare sind derzeit im Volontariat, über sechzig Personen, viele aus Rundfunkanstalten, waren der Einladung gefolgt und konnten wertvolle Einblicke in die angewandte informationswissenschaftliche Forschung und die Medienpraxis gewinnen. Professor Jakob gelang es dabei immer wieder die Relevanz der einzelnen Beiträge für die berufliche Praxis in der Dokumentation auf den Punkt zu bringen.

Werkzeuge der Sprachtechnologie

Zu Beginn gab Melanie Siegel, Professorin für Semantik an der Hochschule Darmstadt, einen Einblick in die Automatische Klassifikation offensiver Sprache und die Erfahrungen aus zwei „shared tasks“. Im Sommer 2019 nahm das Thema „Technik gegen Hass“ in den Medien großen Raum ein und auch die Bundesregierung verabschiedete in einem Kabinettsbeschluss „Neun Punkte gegen den Hass“, darunter eine Meldepflicht für Hass im Netz, Cyber-Stalking, Hetze und aggressive Beleidigung und mehr Schutz für Kommunalpolitiker. „Hassrede“ oder „Hate Speech“ sind in der öffentlichen Diskussion gebrauchte Ausdrücke, denen eine genaue Definition fehlt, wie sie für die Wissenschaft unabdingbar ist. Diese nutzt daher lieber die wertneutrale Bezeichnung „offensive Sprache“.

Begonnen hatte die Forschung mit einer Ausgangsfrage der Interest Group on German Sentiment Analysis

(IGGSA), welche sinnvollen Einsatzgebiete es für die in- zwischen entwickelten Sprachtechnologien gebe. Unter der Leitung von Melanie Siegel und Michael Wiegand, seinerzeit Universität Heidelberg, entschied man sich, offensive aggressive Sprache im Deutschen zu analysieren und rief einen zweistufigen Programmierwettbewerb (shared tasks) aus, ein in der Sprachverarbeitung sehr populäres Vorgehen. Dabei werden Daten zur Verfügung gestellt und die Forschung aufgefordert, mit den Daten zu arbeiten, sie automatisch zu klassifizieren und darauf Analysen durchzuführen. In einer zweiten Runde werden mit den Erfahrungen aus der ersten, Verbesserungen abgeleitet. Die letzte der shared tasks fand Anfang Oktober 2019 statt.

Grundlage sind also Daten, die klassifiziert sein müssen. Die Entscheidung, welche Daten genutzt werden sollten, fiel zugunsten Twitter, weil die Daten im Unterschied zu Facebook öffentlich sind und es eine Programmierschnittstelle gibt. Zunächst wurden mit Stichwörtern „Timelines“ bzw. „User“ identifiziert und davon jeweils die ersten 200 Tweets ausgewählt. Die Daten wurden gesäubert, indem Retweets und kurze Tweets eliminiert wurden. In einem zeitaufwendigen Prozess wurden dann jeweils zu dritt mehrere hundert Tweets intellektuell annotiert und neun Seiten Guidelines erarbeitet, bis eine gute Übereinstimmung bei der Indexierung bestand. Anhand der erarbeiteten Richtlinien wurden anschließend insgesamt 8000 deutschsprachige Tweets annotiert. Zunächst wurde unterschieden, ob es sich darin um offensive Sprache handelt oder nicht und die als offensiven eingestuften Tweets wurden anschließend je nach ihrer Art klassifiziert (Benutzung von Schimpfwörtern, Beleidigung, Diskriminierung, Entmenslichung, Drohung). Um Verzerrungen zu vermeiden und das gesamte politische Spektrum abzubilden wurden zusätzliche Tweets aufgenommen.

Die am Wettbewerb teilnehmenden Forschergruppen erhielten als Trainingsset zunächst Testdaten mit den intellektuell hinzugefügten Annotationen, um ihre Systeme zu trainieren, und anschließend ein Testset, also Testdaten ohne Annotationen, um die Leistungsfähigkeit ihrer Technologie zu demonstrieren. In einem anschließenden

*Kontaktperson: Marlies Ockenfeld, Deutsche Gesellschaft für Information und Wissen e.V., Windmühlstraße 3, 60329 Frankfurt am Main, E-Mail: ockenfeld@dgi-info.de

Workshop präsentieren und erörtern sie dann ihre Ergebnisse. Insgesamt nahmen an dem internationalen Wettbewerb 20 Forschergruppen aus Universitäten und Forschungseinrichtungen teil. Der Tagungsband ist öffentlich.

Beim ersten Wettbewerb 2018 siegte das Programm eines Bachelorstudenten der TU Wien. Die Kategorisierung in offensive oder nicht offensive Sprache wurde dabei mit herkömmlichen Standardverfahren des maschinellen Lernens (Support-Vector Machine, SVM) vorgenommen. Eingesetzt wurden Wörterbücher und Lexika (öffentlich verfügbare Listen, aus dem Englischen übersetzte Listen, eigene Listen, Sentiment-Wörter, Schimpfwörter). Diese Listen stehen der Forschungsgemeinschaft zur Verfügung. Zwischen 70 und 80 Prozent der Tweets ließen sich korrekt zuordnen. Die gewonnene Erkenntnis: Mit guten Wortlisten, aber auch Buchstabenketten (character n-grams) kommt man weit. Implizit offensive Sprache lässt sich allerdings nicht erkennen.

Für den letzten Wettbewerb im Jahr 2019 wurden 2890 offensive Tweets aus dem Wettbewerb von 2018 zusätzlich als explizit/implizit annotiert. 14 Forschergruppen beteiligten sich. Eine rumänische Gruppe gewann mit drei verschiedenen Systemen klar bei der Klassifikation in offensiv/other und ebenfalls bei der Feinklassifikation nach Art der Herabwürdigung. Bei der Klassifikation nach implizit/explicit gewann eine Gruppe aus Potsdam/Düsseldorf/Passau, Bukarest belegte den 3. Platz

Überraschend war, dass Standard Deep-Learning-Methoden nicht unter den Top-Systemen waren, während SVM ebenso wie Word-Embedding und N-Grams weiter ganz oben standen. Trotz der viel größeren Datenmengen erzielte das beste System von 2018 kaum veränderte Erkennungsraten.

Derzeit ist kein neuer Wettbewerb geplant. Es wird jedoch ein Special Issue on offensive Language des JLCL geben, Doktorarbeiten am Forschungszentrum für Angewandte Informatik der Hochschule Darmstadt werden vorbereitet und Forschungsanträge zusammen mit Medienunternehmen sind geplant.

Die vorgestellten Verfahren könnten ein Angebot der Dokumentation für die Redaktionen zur Überwachung der Reaktionen/Kommentare auf ihren Social-Media-Plattformen sein. Sie begründen kein Plädoyer für automatische Filterung, aber die Systeme können Hinweise geben und die Entscheidung über einzuleitende Maßnahmen (Löschen, Antworten, Anzeige erstatten, Statistiken aufstellen, welche Themen zu einem bestimmten Zeitpunkt bestimmend sind) unterstützen.

Je besser die Daten sind, umso besser könnte ein System werden. Ein grundsätzliches Problem ist es, dass die

öffentlichen Daten in Twitter bereits gefiltert sind, interessant für die Sprachanalyse wären die herausgefilterten Tweets. Aus wissenschaftlicher Sicht wäre es interessant, die als Verstöße gemeldeten Beiträge zu bekommen.

Haftung und Lizenzierung

RA Peter Wiechmann, Justiziar des SWR, Mainz, erläuterte die Konsequenzen der EU-Urheberrechtsreform 2019 und ihre Auswirkungen auf die publizistische Verwendung von Informationsressourcen. Im Mittelpunkt seiner Ausführungen standen die beiden neuen Richtlinien zum Urheberrecht in der EU, sozusagen Gerüste, anhand derer die nationalen Gesetzgeber tätig werden müssen, um diese Soll-Vorschrift der EU umzusetzen.

Die erste Richtlinie, Online-SatCab-RL, regelt die Weiterverbreitung von Fernseh- und Hörfunkprogrammen, vor allem die technologieneutrale Weiterentwicklung zur Weitersendung von Hörfunk- und Fernsehprogrammen. Das Ursprungslandprinzips soll sich auch auf Online-Angebote der Sender erstrecken. Ein Rechteerwerb soll nur für das Ursprungsland erforderlich sein, nicht für alle Territorien, in denen das Online-Angebot abrufbar ist, so dass die Notwendigkeit für Geo-Blocking für Hörfunkprogramme entfällt. Geo-Blocking bleibt möglich für reine Online-Angebote und für Sportübertragungen, die vertragliche Vereinbarung einer Pflicht zum Geo-Blocking bleibt möglich. Aus Sicht der Sender ist dies eher eine Verschlimmbesserung, denn die Regelung betrifft nur Sachverhalte, bei denen in der Praxis kein Regelungsbedarf bestand, bei denen jetzt aber Forderungen nach Geo-Blocking oder höherer Vergütung aufkommen könnten. Für die Verbreitung an sich frei zugänglicher Angebote im Ausland könnten Schranken aufgebaut werden.

Größere öffentliche Aufmerksamkeit hat, nicht zuletzt durch den massiven Einsatz von Influencern, die zweite Richtlinie erfahren. Die DSM-RL regelt Leistungsschutzrecht für Verlagsprodukte, die Plattformhaftung für urheberrechtlich geschützte Inhalte, die auf sie hochgeladen werden, das Urhebervertragsrecht und sie enthält Regelungen für Kulturerbe-Einrichtungen.

Nachdem 2002 die Umsetzung der Urheberrechtsrichtlinie vom 22.5.2001 abgeschlossen war, kündigte die Kommission der Europäischen Union 2010 eine „Digitale Agenda für Europa“ zur Klärung der grenzüberschreitenden Lizenzierung von Urheberrechten an. 2012 entschied der EuGH, dass die Betreiber von Social Networks Beiträge nicht mittels automatischer Upload-Filter löschen dürfen. 2018 gab es schließlich einen Richtlinienentwurf Schutz von Pressepublikationen. Ursache war der Value Gap zwischen Vergütung der Urheber und dem Profit der Online-Plattformen, die deren Werke zugänglich machen. Geregelt werden sollten u. a. die Beachtung der Ausschließlichkeitsrechte der Rechteinhaber sowie die Ermöglichung von Text- und Data-Mining.

Die DSM-Richtlinie enthält eine neue Schrankenregelung (Text und Datamining zum Zweck der wissenschaftlichen Forschung), die in Deutschland bereits in § 60 UrhG berücksichtigt ist (Digitale Nutzung von Werken zu Unterrichtszwecken und Vervielfältigung von Werken durch Kulturerbe-Einrichtungen zur Bewahrung ihrer Bestände). Ferner enthält sie Regelungen zur Verbesserung der Lizenzierungspraxis, indem für die Lizenzierung vergriffener Werke (§§ 51 ff. VGG) eine Extended Collective Licence eingeführt wird, die für Werke von Urhebern gilt, die nicht Mitglied einer Verwertungsgesellschaft sind. Neu eingeführt werden auch der Einsatz von Mediatoren für Schwierigkeiten bei der Lizenzierung audiovisueller Werke für Videoabrufdienste und der Schutz für Fotos von gemeinfreien Werken der bildenden Kunst soll entfallen.

Hinsichtlich der Regelungen zur Schaffung eines funktionsfähigen Marktes für Urheberrechtsschutz ist bis auf die Lizenzierung und Haftung für Online-Plattformen in Deutschland alles schon im UrhG geregelt. Abweichungen gibt es im Detail bei der Auslegung dessen, was als Einrichtung des kulturellen Erbes gilt, vor allem, ob Rundfunkanstalten bzw. ihre Archive dazu gehören oder nicht, sowie bei der Unterrichts- und Lehrtätigkeit, wo es eine Auseinandersetzung gibt, ob das Abspielen von mehr als zehn Minuten Schulfunksendung lizenzfrei ist oder nicht. Wenn man zu Einrichtungen des Kulturerbes zählt, hat man den Vorteil erleichterter Lizenzierungsbestimmungen oder das Recht zu Text- und Datamining.

Hinsichtlich der Erhaltung des Kulturerbes sind die Regelungen in § 60 UrhG detaillierter und weitergehend. Erlaubt ist die Zugänglichmachung an Terminals in den Räumlichkeiten der Einrichtung; der Forderung, dass diese Werke auch generell öffentlich zugänglich gemacht werden dürfen, hat der EU-Gesetzgeber nur für vergriffene Werke entsprochen. Als vergriffen gilt ein Werk, wenn davon ausgegangen werden kann, dass das gesamte Werk auf den üblichen Vertriebswegen für die Öffentlichkeit

nicht erhältlich ist. Zum Nachweis vergriffener Werke gibt es ein Register beim DPMA, in das inzwischen 26.671 Werke, überwiegend Bücher, eingetragen sind. Archive ließen sich zwar für die allgemeine Öffentlichkeit öffnen, doch ist die Lizenzierung meistens sehr teuer.

Für gemeinfreie Werke der bildenden Kunst soll gelten, dass nach Ablauf der Schutzdauer eines Werkes der bildenden Kunst z.B. Fotos der Werke weder urheberrechtlich noch durch verwandte Schutzrechte geschützt sind, es sei denn, die Fotos stellen eine eigene geistige Schöpfung dar. Laut der BGH-Entscheidung „Museumsfoto“ können selbst zweidimensionale Reproduktionen gemeinfreier Werke verboten sein.

Beim Schutz für Presseveröffentlichungen wird eine Haftung der Plattformen für das Teilen von Online-Inhalten eingeführt. Facebook und Youtube werden nicht genannt, sondern durch abstrakte Formulierungen umschrieben: Auch der Betreiber einer Plattform für das Teilen von Online-Inhalten ist urheberrechtlich verantwortlich für die Zugänglichmachung von Inhalten, die seine Nutzer hochladen. Der Betreiber muss daher eine Erlaubnis des Rechteinhabers einholen (Lizenzpflicht des Plattformbetreibers). Ohne Erlaubnis haftet der Plattformbetreiber urheberrechtlich, es sei denn er weist nach, dass er alle Anstrengungen unternommen hat, um die Erlaubnis einzuholen. Nach entsprechenden Hinweisen durch den Rechteinhaber muss er löschen. Hier besteht die Gefahr der Zensur. Es wird befürchtet, dass das Urheberrecht immer öfter als Mittel zur Unterdrückung unerwünschter Kritik missbraucht werden könnte. Upload-Filter können nicht erkennen, ob die Nutzung aufgrund urheberrechtlicher Schranken erlaubt ist oder nicht. Es entsteht der Eindruck, dass vor allem die Interessen der Verwerter, aber nicht die der Urheber geschützt werden.

Das in der Richtlinie regelte Urhebervertragsrecht schreibt einen Anspruch auf angemessene Vergütung fest sowie eine Transparenzpflicht (einmal jährlich unaufgefordert Benachrichtigung über Art der Verwertung, erzielte Einnahmen, Forderungen gegenüber Lizenznehmer).

2018 entstand vor allem in Deutschland ein großer öffentlicher Widerstand gegen die Richtlinie, sodass sich die Bundesregierung veranlasst sah bei ihrer Zustimmung zur Richtlinie eine Protokollerklärung abzugeben, dass es keinen Einsatz von Upload-Filter geben solle, die Regelung auf Facebook und Youtube zielt und Wikipedia, Hochschulen, Blogs, Foren, Github, Messengerdienste, Verkaufsportale und Cloud-Dienste ausgenommen seien, Zitatfreiheit garantiert sein müsse und Schutz und Vergütung für Kreative unbedingt notwendig seien.

Forschungsdaten in den Geistes- und Kulturwissenschaften

Prof. Dr. Stefan Schmunk, Professor für Digitale Bibliotheken an der Hochschule Darmstadt, gab einen Überblick über den Aufbau einer Infrastruktur für die Generierung, Erschließung, Speicherung und Langzeitarchivierung von Forschungsdaten in den Geistes- und Kulturwissenschaften und gab einen Einblick in technische Lösungen für Interoperabilität. Kernbereiche des von der DFG geförderten digitalen Wandels in Wissenschaft und Forschung sind u. a. die Verwendung und Analyse von Forschungsdaten sowie neue Publikations- und Präsentationsmöglichkeiten. Die Allianz der deutschen Wissenschaftsorganisationen hat in ihrem Leitbild 2018 bis 2022 die drei Schwerpunkte Open Access – Open Science – Forschungsdaten gesetzt. Als ein Beispiel digitaler Transformation in den Geisteswissenschaften, aber auch für das gemeinsame Arbeiten präsentierte er die Faust-Edition www.faustedition.net. Die Bibliotheken erleben einen Wandel. Sie sollen digitale Zugänglichkeit gewährleisten und müssen sich deshalb damit auseinandersetzen, wie Daten erhoben und dokumentiert werden.

Zwei große Initiativen sind TextGrid (2006 gestartet) und das BMBF-geförderte DARIAH-DE (nach acht Jahren Entwicklung seit 2019 im Dauerbetrieb). DARIAH-DE wird von einem Konsortialverbund aus Universitäten, Forschungszentren, KMU, NGO, Bibliotheken u. a. entwickelt und betrieben. Über beide Vorhaben wurde in dieser Zeitschrift bereits berichtet.

Digitale geistes- und kulturwissenschaftliche Forschungsdaten innerhalb von DARIAH-DE sind definiert als all jene Quellen/Materialien und Ergebnisse, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und in maschinenlesbarer Form zum Zwecke der Archivierung, Zitierbarkeit und zur weiteren Verarbeitung aufbewahrt werden können.

Eine Forschungsinfrastruktur, so betonte Schmunk, besteht vor allem aus Menschen und Kompetenzen. Auf der technischen Ebene sind dafür kollaborative Arbeitsumgebungen, PID Services für persistente Identifizierung, Speicherungs- und Archivierungssysteme mit Zugriffsmanagement und Datenbanken sowie möglichst standardisierte Metadaten unabdingbar.

Wie arbeitsintensiv und zeitaufwendig Digitalisierungsvorhaben in den Kulturwissenschaften sein können wird z. B. am Projekt Theodor Fontane's Notizbücher deutlich, das inzwischen im neunten Jahr läuft (fontane-nb-dariah.eu/index.html).

Etwa die Hälfte des Kulturguts vom 16. bis 18. Jahrhundert ist inzwischen digitalisiert. Dabei gibt es unterschiedliche Digitalisierungsklassen, mit der die Erschließungstiefen für die wissenschaftliche Nutzung unterschieden werden können. Jeder kann im DARIAH-DE Collection Registry eine eigene Sammlung anlegen. Allerdings besteht eine große Diskrepanz zwischen der Bereitschaft Daten anderer zu nutzen und eigene Daten zur Verfügung zu stellen. Um Forscher, die eigene Daten hochladen wollen, nicht durch komplexe Datenerfassungsschemata mit vielen Pflichtfeldern abzuschrecken, hat man sich auf die 15 Pflichtfelder von Dublin Core beschränkt. Die Migration unterschiedlicher Formate und die Datenpflege sind bisher ungelöst, auch wenn es maschinelles Mapping in einigen Bereichen bereits gibt. Festzuhalten ist, dass die Dokumentation von Prozessen und Vorgehensweisen an fachkundigen Menschen, wie Dokumentaren, hängt. Die Pflege ist auf Dauer genauso teuer wie Entwicklung. Auch die Akzeptanz ist bisher sehr überschaubar. Am selbstorganisierten Management der eigenen Publikationen und Daten über ORCID bei DARIAH-DE machen bisher nur drei Wissenschaftler mit.

Triple-Stores im Dokumentationsprozess

Professor Geribert Jakob stellte Möglichkeiten der semantischen Recherche in Ontologie-gestützten Dokumentationssystemen vor und skizzierte die dafür erforderlichen Vorarbeiten von Dokumentaren beim Aufbau der Triple-Stores. Die semantische Recherche ist eine Cross-Media-Erweiterung für traditionelles Retrieval in Medienarchiven. Der Standard RDF (Ressource Description Framework) definiert wie Aussagen in der Form Subjekt – Prädikat – Objekt maschinell gespeichert werden. Subjekte und Objekte werden durch eine URI repräsentiert, die auf eine dauerhafte repräsentative Informations- oder Datenquelle verlinkt. Prädikate sind Informationsressourcen, die eine semantische Beziehung beschreiben und deren Definition einer frei zugänglichen Ontologie entstammt, z. B. www.onto-name.org.

RDF-Triple werden als Datensätze tabellarisch notiert und ergeben einen Triple-Store, eine dreispaltige Tabelle. Triple-Stores, die nur formale und linguistische Prädikate enthalten und sich auf die Begriffsbeziehung „besitzt genormte Eigenschaft“ reduzieren, sind identisch mit Thesauri, die letztlich nur das Wiederauffinden entsprechend beschriebener und indexierter und klassierter Informationsressourcen erlauben. Semantische

Prädikate, die für die vernetzte Beschreibung in keiner bestehenden und zugänglichen Ontologie auffindbar sind, müssen in einer eigenen dafür zu erarbeitenden Ontologie definiert werden. Inzwischen gibt es ca. tausend Ontologien.

Nicht webfähige Zugriffe werden mit einem SQL-Wrapper eindeutig adressierbar gemacht. Die SQL-Abfrage wird einer eindeutigen URI zugeordnet, so dass solche Ressourcen in einen Triple-Store eingetragen werden.

Ontologische Suche basierend auf Wikidata werden zum Beispiel von Google und Wolfram Alpha angeboten.

Die Stärke von Dokumentaren ist es, die Semantik der Wissensbereiche zu liefern, für die sie zuständig sind. Sie spielen deshalb bei der Befüllung von Triple-Stores eine wichtige Rolle. Es gilt existierende Datenbestände zu kennen und soweit wie möglich zu nutzen. Mit der semantischen Suche wird die Recherche über den kontrollierten Essenzbestand hinaus durch Zugriff auf Web-Ressourcen erweitert, auf Ontologien wie DBpedia oder foaf. Die externen Quellen liefern semantische Beziehungen. Neben den gefundenen relevanten Prädikaten müssen all solche identifiziert werden, die für den internen Essenzbestand relevant sind, um alle semantischen Beziehungen zwischen den Essenzen beschreiben zu können. Die Abfragen mit dem Ziel einschlägige Dokumente zu finden, wandeln sich zu den bekannten journalistischen Fragen wer, wo, wann, wie, wieviel, warum und deren Kombination. Die Antworten der Systeme wandeln sich von nachzubearbeitenden Ergebnis-/Trefferlisten zu Antworten auf Fragen, die relevante Antworten liefern. Dokumentare bestimmen die einschlägigen externen Ontologien für vorhandene Objekte und Subjekte in den bestehenden Dokumentationsspeichern, die relevanten Prädikate in externen Ontologien sowie die fehlenden semantischen Beziehungen und bauen hierfür eine interne Ontologie auf. Sie sind die Know-how-Träger für die Triple-Erzeugung für alle Objekte und Subjekte in den bestehenden Dokumentationsspeichern und Medienarchiven. Der Aufwand für Aufbereitung der Inhalte eines Medienarchivs für die semantische Suche kann das 20- bis 100fache des IT-Entwicklungsaufwands betragen und ist eine längerfristige Zukunftsaufgabe.

Metadatenmanagement und Genealogie

Stefan Doğanay, Abteilungsleiter Contentservices bei der Infonetwork GmbH, gab einen Einblick in das Metadatenmanagement und die Material-Genealogie bei der Medien-

gruppe RTL Deutschland. 750 Mitarbeiter in Köln, Berlin, München, London, Paris, New York, Washington und Tel Aviv produzieren journalistische Inhalte für 17 Redaktionen im linearen TV und für Online-Dienste. Der tägliche Video Ingest umfasst 300 bis 350 Stunden in ca. 2500 Dateien, der Kosten und Effizienzdruck ist hoch.

Zentrales Steuerungscockpit ist das Redaktionsportal, das auch von den einzelnen Redakteuren für den Selbstschnitt genutzt wird. Mit einem Metadaten-Enricher werden Daten an KI-Anbieter geschickt, um die Indexierung anzureichern. Geplant sind eine Gesichtserkennung bei Bundestagsdebatten zur Verschlagwortung der O-Töne sowie deren automatisierte Verschriftlichung.

Die Metadaten beziehen sich immer auf einzelne Shots. Bei der Metadaten-Vererbung werden nur Referenzen weitergegeben, die am Ende wieder zusammengefügt werden. Neu geschnittenes Material muss dadurch nicht neu erschlossen werden, denn es werden die Metadaten des Ausgangsmaterials übernommen. Jeder hat Zugriff auf die Schnittlisten. Jede Sequenz hat eine ID und einen Time Code, zu jedem Shot sind die Lizenzinformationen gespeichert. Es wird alles innerhalb des Systems produziert, wenn extern zugearbeitet wird, sind die produzierten Beiträge häufig nicht mit guten Metadaten versehen.

Um die Treffermüllmenge zu verringern, vererbt das Archivmaterial keine Metadaten.

Alle Beiträge werden stets auf das Rohmaterial zurückgeführt, aus dem sie zusammengeschnitten worden sind. Dazu wird eine Graphdatenbank zum Media Graph Traversal eingesetzt. Unter Überspringen sämtlicher zwischenzeitlich produzierter MAZ wird immer wieder auf das Rohmaterial verlinkt. Jedes MAZ hat dabei Verbindung zu allen verwendeten Rohmaterialien.

Dies bietet Vorteile für verschiedene Anwendungsfälle. Wenn Material gesperrt werden muss, sei es aus lizenzrechtlichen Gründen oder weil das Einverständnis zurückgezogen worden ist, lässt sich im Redaktionssystem durch Rücksprung auf das Rohmaterial feststellen, in welchen Beiträgen das Material überall genutzt worden ist. Wenn geklärt werden soll, ob eine Erstverwendung erfolgt ist, kann durch eine Vorwärts-Traversal nach der Nutzung des Materials recherchiert werden und dann in den Metadaten bei Lizenzen eine entsprechende Eintragung erfolgen. Kontingentverträge, z. B. mit Getty, erlauben die Nutzung innerhalb eines begrenzten Zeitraums. Dies lässt sich ebenfalls im System gut verwalten.

Geplant ist, künftig auch Rohmaterial aus dem Materialcluster in Form einer Empfehlung für einen neuen Beitrag herauszufiltern. Die Redakteure stehen unter enormem Zeitdruck und schätzen daher die Vorarbeit der Dokumentation.

Mithilfe der Graphen sind auch Business Intelligence-Anwendungen möglich, indem sich Überschneidungen in Sendungen herausfinden lassen oder von welchen Lieferanten/Agenturen besonders viel Material genutzt wird.

Grundsätzlich wird nach dem Grundsatz verfahren „lieber keine Metadaten als falsche Metadaten“. Bei Downloads oder Ingest-Prozessen liegen Metadaten in der Regel vor, für anderen Materialien gelten Prioritätsregeln für verschiedene Themen und Wiederverwendungspotenziale. Bei zugelieferten Beiträgen erfolgt die Einordnung in eine Kategorie durch eine KI.

Trau – schau – wem

Jörn Ratering vom Programmarchiv des ZDF gab einen eindrucksvollen Ausblick auf neueste Verfahren der *Informationsforensik* und *Webquellenvalidierung* und zeigte damit noch einmal ein spannendes Arbeitsgebiet für Dokumentare auf.

Er konzentrierte sich dabei auf drei aktuelle Entwicklungen zur Verifizierung von Informationen. Zunächst erläuterte er am Beispiel der Familie, die im niederländischen Ruinerwold im Keller gefangen gehalten wurde, wie viele Informationen sich aus Social Networks gewinnen lassen und welche Widersprüche und Fälschungen sich durch eine genaue Betrachtung von Fotos und Posts aufdecken lassen. Das angebliche Foto des von der Boulevard-Presse als Sekten-Josef bezeichneten Familienvaters ist bei CanStock für sieben US\$ erhältlich. Auf der Facebook-Seite von Jan, dem jungen Mann, der in einer Bar Hilfe gesucht hatte, findet sich eine Eintragung, die mit dem Datum 1. August 2010 versehen ist, in der er angibt, jetzt nach Ruinerwold gezogen zu sei. Tatsächlich ist diese Eintragung erst am 15. September 2019 auf der Facebook-Seite veröffentlicht worden, wie sich nachweisen ließ.

Danach ging es um derzeitige Entwicklungen von Werkzeugen zur Verifikation von Informationen. Mit der

Rückwärtssuchmaschine yandex lassen sich Fotos als Montagen entlarven und feststellen, aus welchen einzelnen Fotos sie zusammengesetzt wurden. Die Suchmaschine arbeitet offenbar zusätzlich auch mit einer Gesichtserkennungs-Software, weil Personen sogar dann richtig erkannt werden, wenn zwar andere, nicht aber genau das gesuchte Foto im Internet existiert.

Abschließend veranschaulichte Ratering an Beispielen aktuelle Neuentwicklungen in Richtung Deepfakes. Mittels künstlicher Intelligenzen lassen sich damit Stimmen und Gesichter nachbilden, so dass man nicht mehr unterscheiden kann, ob es sich um den richtigen Donald Trump oder den falschen handelt. Doch gibt es umgekehrt auch Software wie Proof Mode von Truepic, die digitale Aufnahmen bei der Entstehung mit einem Verifizierungscodier versieht, sodass Verfälschungen leicht aufgedeckt werden können.

Insgesamt vermittelte die gut organisierte Veranstaltung den diskussionsfreudigen Teilnehmerinnen und Teilnehmern einen gelungenen Einblick in die vielfältige und anspruchsvolle Arbeitswelt der in der Dokumentation tätigen Informationsfachleute und in neue interessante durch die informationswissenschaftliche Forschung eröffnete Arbeitsgebiete.

Deskriptoren: Tagung, Mediendokumentation, Sprachtechnologie, Forschungsdaten, Geisteswissenschaften, Recht, Urheberrecht, Verifikation, Fälschung, Hochschule Darmstadt

Marlies Ockenfeld

Deutsche Gesellschaft für Information und Wissen e.V.
Windmühlstraße 3
60329 Frankfurt am Main
ockenfeld@dgi-info.de