FLEXIBLE DIGITALE KURATIERUNGSTECHNOLOGIEN FÜR VERSCHIEDENE BRANCHEN UND ANWENDUNGSSZENARIEN

Georg Rehm

Forschungsbereich Sprachtechnologie, DFKI GmbH, Deutschland,
georg.rehm@dfki.de

KURZDARSTELLUNG: Der Beitrag stellt das BMBF-geförderte Verbundprojekt „Digitale Kuratierungstechnologien" vor, in dem das DFKI gemeinsam mit vier in Berlin ansässigen KMU-Partnern der Frage nachgeht, inwiefern semantische Sprach- und Wissenstechnologien eingesetzt werden können, um die branchenspezifischen Bedarfe der bei den vier KMU-Partnern tätigen Wissensarbeiterinnen und Wissensarbeiter zu unterstützen, beispielsweise hinsichtlich Effizienz, Qualität und Abdeckungsgrad der entstehenden Contents und Informationsprodukte.

1. EINFÜHRUNG

Das Kuratieren digitaler Informationen, Daten, Meldungen und Medieninhalte hat sich in den vergangenen Jahren als eine grundlegende Tätigkeit mit neuen Anforderungen herauskristallisiert, die von handelsüblichen Content-Management-Systemen schon längst nicht mehr abgedeckt werden. Abstrakt formuliert ist Kuratierung ein komplexer zeit- und wissensintensiver Prozess, in dem Experten – z.B. Redakteure, Wissenschaftler oder interdisziplinäre, verteilte Teams – aus einer thematisch typischerweise homogenen, oft aber auch heterogenen Sammlung von Quellen ein neues, in sich kohärentes und abgestimmtes Gesamtwerk entwickeln, das auf einen spezifischen Fokus ausgerichtet ist, also eine spezielle kommunikative Funktion besitzt.

Die erforderlichen Arbeiten umfassen das Auswählen, Abstrahieren, Einordnen, Internationalisieren, Zusammenfassen, Anreichern, Sortieren, Visualisieren und das zusätzliche Erläutern, Umschreiben, Neuformulieren und Ergänzen der Inhalte, wobei insbesondere zu berücksichtigen ist, dass Geschwindigkeit, Volumen und Anzahl der Quellen der zu verarbeitenden Informationen stetig wachsen (im Digitalkontext z.B. Online-Zeitungen, Nachrichtenportale, Fachinformationen, aber natürlich auch die sozialen Netzwerke wie z.B. Twitter, Facebook, Instagram, Pinterest etc.).

Ein Beispiel: Die Entwicklung eines für ein Museum vorgesehenen interaktiven Exponats, das bei Ausgrabungen entdeckte Artefakte mit Fotos, Beschreibungen und Zeitangaben auf einer interaktiven Karte visualisiert, erfordert die Auswahl der geeigneten Objekte, die Erstellung entsprechender Inhalte (Beschreibungen, Fotos, Videos etc.), die Gestaltung der Karte, die Festlegung thematischer Perspektiven sowie natürlich auch die eigentliche Implementierung. Ein zweites Beispiel aus dem Bereich Online: Die Erstellung eines deutschsprachigen Artikels über eine Naturkatastrophe in Süd-Ost-Asien erfordert die Durchsicht von Agenturmeldungen (Lesen, Sortieren, Auswählen), die Recherche in diversen sozialen Netzwerken (Auswahl und Übersetzung von Texten, Zitaten, evtl. auch Bild- und Videomaterial, Sicherstellung der Authentizität etc.) sowie auch in wissenschaftlichen Diskussionsforen (etwa zum Thema Klimawandel, Tsunamiforschung etc.) und die anschließende Zusammenfügung der gefundenen Bausteine zu einem neuen, in sich geschlossenen Beitrag, der evtl. zu späteren Zeitpunkten fortlaufend ergänzt wird.

Typische Arbeitsabläufe dieser Form lassen sich in zahlreichen Branchen und Domänen identifizieren, in denen – bewusst sehr abstrakt formuliert – eine oder mehrere Personen aus eingehenden Informationen ausgehende Informationen produzieren. Nicht nur im Rahmen der aktuellen Diskussion zu den Themen „Digitalisierung" und „Neue Arbeit" stellt sich in diesem Zusammenhang die Frage, wie derartige Workflows in unterschiedlichen

Bereichen und Branchen bestmöglich durch smarte semantische Technologien unterstützt werden können.

2. KURATIERUNGSTECHNOLOGIEN

Zunächst gilt es zu erläutern, was digitale Kuratierungstechnologien eigentlich sind. Die entsprechenden Grundlagen werden derzeit im Rahmen eines vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Verbundprojekts untersucht, in dem das DFKI als Forschungspartner und Koordinator gemeinsam mit den vier Berliner Unternehmen ART+COM AG, Condat AG, 3pc GmbH und Kreuzwerker GmbH „Digitale Kuratierungstechnologien" (DKT) in Form diverser Prototypen entwickelt (siehe auch http://www.digitale-kuratierung.de).
Das Ziel dieses ersten Pilotprojekts ist es, komplexe, von Redakteuren und Wissensarbeitern durchgeführte digitale Kuratierungsprozesse durch Sprach- und Wissenstechnologien zu unterstützen. Das DFKI bringt verschiedene Komponenten aus diesem Bereich ein, entwickelt diese weiter und baut gemeinsam mit den o.g. KMU-Partnern eine Plattform für digitale Kuratierungstechnologien auf, die u.a. Funktionen zur Recherche, Anreicherung, Analyse, Kombination (z.B. thematisch, chronologisch, räumlich), Zusammenfassung und Internationalisierung von Inhalten zur Verfügung stellt (siehe Abb. 1). Die einzelnen Kuratierungstechnologien werden als RESTful APIs angeboten, die flexibel zu Workflows von Services kombiniert werden können. Ermöglicht wird dies über eine flexible Plattform, die intern mit Annotationen im Natural Language Processing Interchange Format (NIF) arbeitet und in dem ebenfalls vom DFKI koordinierten EU-Projekt FREME entwickelt wurde (mehr Informationen hierzu unter http://www.freme-project.eu).
Die über diese Plattform zur Verfügung stehenden Kuratierungsservices können von den vier KMU-Partnern des Verbundes im Rahmen ihrer jeweiligen Nutzungsszenarien (nahezu) beliebig in die jeweiligen branchenspezifischen Lösungen integriert werden, was wiederum die Implementierung branchenspezifischer Workflows und skalierbarer Anwendungen ermöglicht. Die Plattform erlaubt es also den Industriepartnern, innovative und effizienz- sowie qualitätssteigernde Lösungen für ihre unterschiedliche Branchen effizienter zu entwickeln, zu betreiben und zu verwerten. In dem Projekt stehen die folgenden vier Branchen im Fokus:

- Museen und Showrooms
- TV-/Radio und Web-TV-Sender
- Verlage und Medienhäuser
- Öffentliche Archive

Die vom DFKI eingebrachten Technologien umfassen computerlinguistische Methoden, Komponenten und Ansätze aus dem Gebiet der Sprach- und Wissenstechnologien, die im Kontext zahlreicher Projekte (BMBF, BMWi, EU/EC etc.) entwickelt wurden. Die Technologien können grob den drei Bereichen *Semantische Analyse* (Informationsextraktion, Named Entity Recognition, Temporale Analyse, Geolokalisierung, Annotation mit allgemeinen Metadaten, Clustering, Klassifikation, Sentiment-Analyse), *Semantische Generierung* (Textgenerierung, Semantic Storytelling) und *Mehrsprachige Technologien* (maschinelle Übersetzung, mehrsprachige Linked Data) zugeordnet werden.
Die Plattform besitzt im Kontext der Wertschöpfungskette eine spezielle Bedeutung. Je nach Anwendungsfall und Branche fällt die Wertschöpfungskette zur Kuratierung von Inhalten unterschiedlich aus, so dass (mindestens) drei Klassen beteiligter Akteure zu unterscheiden sind:

- die kuratierende Institution, z.B. Museum, Fernsehsender, Verlag oder Archiv;
- Dienstleister/Agenturen, die für die kuratierende Institution Inhalte und Technologien bereitstellen bzw. Komplettlösungen entwickeln;
- an der Kuratierung beteiligte Redakteure und Wissensarbeiter, z.B. interne Mitarbeiter oder Dienstleister, aber auch externe Wissenschaftler, Experten oder Freiberufler.

Die Plattform für digitale Kuratierungstechnologien soll diesen Akteuren umfassende Funktionalitäten bieten, die möglichst den gesamten Kuratierungsprozess flexibel unterstützen. Durch den Einsatz von Sprach- und Wissenstechnologien können einzelne, bisher typischerweise rein manuell bzw. intellektuell durchgeführte Kuratierungstätigkeiten zumindest (teil-)automatisiert werden. Die Nutzer der Plattform

können größere Mengen an Inhalten schneller sichten und weiterverarbeiten. Insgesamt wird mit der Plattform somit eine Effizienzsteigerung und Kostensenkung des Kuratierungsprozesses angestrebt – bei gleichbleibender oder sogar verbesserter Qualität des erzeugten Outputs.

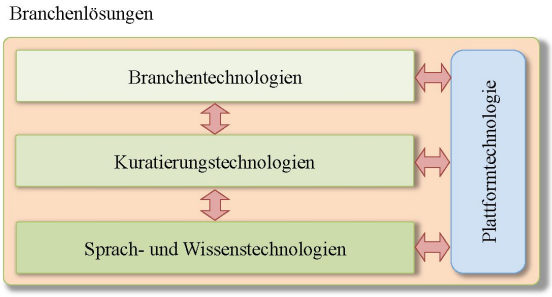

Abb. 1: *Die einzelnen Schichten der Plattform für Digitale Kuratierungstechnologien*

Das DFKI entwickelt die vorhandenen bzw. verfügbaren Komponenten so weiter, dass sie den Anforderungen der KMU-Partner genügen und im Rahmen erster prototypischer Anwendungen evaluiert werden können; die tatsächliche Integration in die jeweiligen Branchenlösungen ist für einen späteren Zeitpunkt geplant. Um in diesem Zusammenhang eine erfolgreiche Markteinführung zu gewährleisten, fokussieren wir insbesondere die folgenden Zielmerkmale der Plattform: Wir streben vollständig integrierte, robuste, performante und skalierbare Komponenten mit flexiblen APIs an, die eine effiziente Einbettung in die branchenspezifische Kuratierungsworkflows erlauben. Ferner ist eine möglichst einfache Nutzbarkeit der Cloud-Plattform von zentraler Bedeutung (SaaS). Für die Branchenlösungen ist auf Seiten der KMU-Partner jeweils eine sehr hohe Usability hinsichtlich User Interfaces, Interaktionsdesign und Informationsvisualisierung vorgesehen. Neben den jeweils eigenständigen und branchenspezifischen Schnittstellen der vier KMU-Partner arbeitet das DFKI an einem grafischen Kuratierungs-Dashboard, dessen aktueller Stand in den Abbildungen 2 und 3 exemplarisch dargestellt wird.
Einen dynamischeren Eindruck vermittelt ein kurzes Screencast-Video, das unter https://www.youtube.com/watch?v=TgP_Txoo buU zur Verfügung steht.

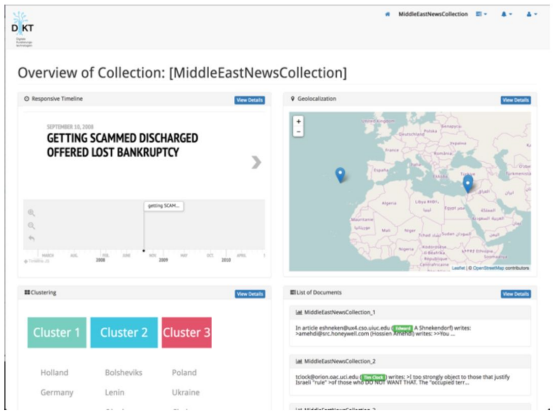

Abb. 2: *Das Kuratierungs-Dashboard, das einen unmittelbaren Zugriff auf die Kuratierungsservices bietet (1/2)*

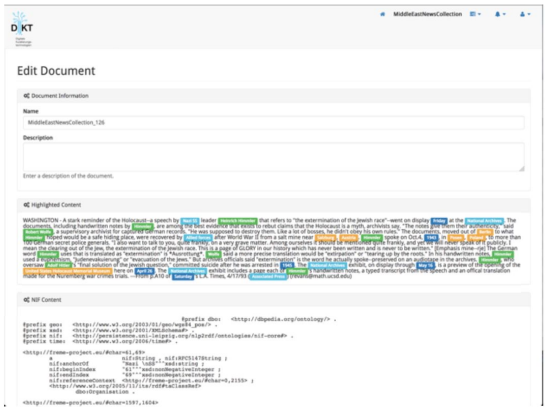

Abb. 3: *Das Kuratierungs-Dashboard, das einen unmittelbaren Zugriff auf die Kuratierungsservices bietet (2/2)*

3. DANKSAGUNG

Das Projekt DKT wird unterstützt durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Programms “Unternehmen Region”, Wachstumskern-Potenzial (Nr. 03WKP45).

4. LITERATURHINWEISE

Die nachfolgenden Literaturhinweise liefern vertiefende Informationen zu spezifischen Aspekten des Projekts DKT.

1. Bourgonje, Peter, Julián Moreno Schneider, Georg Rehm und Felix Sasaki. „Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows". In: Aldo Gangemi und Claire Gardent, Hrsg., Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016), S. 13-16, Edinburgh, UK, September 2016. The Association for Computational Linguistics.

2. Bourgonje, Peter, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki und Ankit Srivastava. „Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer." In: Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer und Christoph Lange, Hrsg., The Semantic Web: ESWC 2016 Satellite Events, June 2016. Im Druck.

3. Moreno Schneider, Julián, Peter Bourgonje, Jan Nehring, Georg Rehm, Felix Sasaki und Ankit Srivastava. „Towards Semantic Story Telling with Digital Curation Technologies." In: Larry Birnbaum, Octavian Popescuk und Carlo Strapparava, Hrsg., Proceedings of Natural Language Processing meets Journalism – IJCAI-16 Workshop (NLPMJ 2016), New York, July 2016.

4. Neudecker, Clemens und Georg Rehm. „Digitale Kuratierungstechnologien für Bibliotheken." Zeitschrift für Bibliothekskultur 027.7, November 2016. Im Druck.

5. Rehm, Georg. „Der Mensch bleibt im Mittelpunkt – Smarte Technologien für alle Branchen." Vitako Aktuell. Zeitschrift der Bundes-Arbeitsgemeinschaft der Kommunalen IT-Dienstleister e.V., 2-2016:26-27, 2016.

6. Rehm, Georg und Felix Sasaki. „Digital Curation Technologies." In: Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016), Riga, Latvia, May 2016. Im Druck.

7. Rehm, Georg und Felix Sasaki. „Digitale Kuratierungstechnologien – Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte." In: Proceedings der Frühjahrstagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL 2015), S. 138-139, Duisburg, 2015. 30. September-2. Oktober 2015.

8. Srivastava, Ankit, Felix Sasaki, Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring und Georg Rehm. „How to Configure Statistical Machine Translation with Linked Open Data Resources". In Proceedings of Translating and the Computer 38, London, UK, November 2016. Im Druck.