ARCHIVE ZUM SPRECHEN BRINGEN – SEMANTIC STORYTELLING ODER DER REDAKTIONSWORKFLOW DER ZUKUNFT

Armin Berger

Geschäftsführer 3pc GmbH Neue Kommunikation, Deutschland, info@3pc.de

KURZDARSTELLUNG: Mit der steigenden Anzahl digitaler Archivalien steigt auch der Bedarf an innovativen Werkzeugen für die Verfügbarmachung im Netz. Gefragt sind Lösungen, die Archivare, Editoren, Redakteure und Autoren dabei unterstützen, ansprechende Publikationen im Netz zu erschaffen. Im Forschungsprojekt DKT – Digitale Kuratierungstechnologien entwickelt 3pc deshalb ein smartes Autorensystem für Semantic Storytelling. Es integriert moderne Sprach- und Wissenstechnologien zu einem neuartigen Redaktionsworkflow, der sowohl die semantische Anreicherung von Archivalien mit semi-automatischen Verfahren unterstützt (NER – Named Entity Recognition), als auch das digitale Storytelling durch intelligente Funktionalitäten (semantische Suche und Empfehlungen) unterstützt. Der Beitrag stellt den aktuellen Stand der Forschungsarbeiten anhand eines Prototypen dar und gibt einen Ausblick auf die zukünftigen Entwicklungen.

1. EINFÜHRUNG

Längst ist in Deutschland die Digitalisierung von Archivalien zur alltäglichen Praxis in den unterschiedlichsten Kultureinrichtungen geworden. In der Vergangenheit drehte sich die Fachdiskussion vor allen Dingen um Fragen zur effizienten Erzeugung und korrekten Archivierung von Digitalisaten. Heute geht es verstärkt um deren sinnvolle Aufbereitung für die Präsentation im Netz. Hier reicht das Spektrum von digitalen Zeitungsbeständen des 19. und 20. Jahrhunderts, über digitale Briefeditionen bis hin zu eher populärwissenschaftlichen Angeboten wie beispielsweise die Online-Mediathek der Stasi-Unterlagenbehörde ("Stasi-Mediathek") [1]. Allen Projekten gemein ist ein hoher manueller Aufwand sowohl im Hinblick auf klassische Aufgaben des Editierens wie die Indexierung (Verschlagwortung), als auch redaktionelle Aufgabenstellungen wie die Erstellung von Themendossiers, Storytelling etc.

Im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekts Digitale Kuratierungstechnologien entwickelt 3pc ein prototypisches Autorensystem für digitales Storytelling [2]. Das System soll seine Anwender dazu befähigen, über eine intuitiv bedienbare Benutzeroberfläche digitale Archivbestände für eine ansprechende Online-Präsentation aufzubereiten.

2. DIGITALISIERUNG ALS HERAUSFORDERUNG UND CHANCE

Bereits seit Jahren ist 3pc im Kulturbereich engagiert und hat in den letzten Jahren eine Reihe komplexer und umfangreicher Web-Projekte für Bibliotheken, Archive und Kultureinrichtungen realisiert. Neben der Konzeption ansprechender Web-Designs und der Entwicklung benutzerfreundlicher Oberflächen stand in diesen Projekten immer auch die Erschließung umfangreicher digitaler Archive im Vordergrund. Die darin abgelegten digitalen Medienobjekte mussten dazu in aufwändiger manueller Arbeit von speziell geschulten Wissensarbeitern aufbereitet werden, um sie anschließend für eine redaktionelle Erschließung benutzbar zu machen.

Schnell wurde deutlich, dass es neuer technischer Hilfsmittel bedarf, um die steigende Anzahl digital verfügbarer Archivalien für eine

ansprechende Präsentation im Netz aufzubereiten – und das bei einem vertretbaren Kosten-Nutzen-Verhältnis. Als Lösung boten sich aktuelle Sprach- und Wissenstechnologien an, zu denen im engeren Sinn auch die semantischen Web-Technologien gehörten [3]. Die Idee des Semantic Storytelling war bei 3pc geboren worden, bevor es folgerichtig zur Entwicklung erster eigener technologischer Lösungsansätze im bereits eingangs erwähnten Forschungs- und Verbundprojekt Digitale Kuratierungs-technolgien kam.

2.1 SEMANTIC STORY TELLING

Inspiriert von den Möglichkeiten des Semantic Web entwickelte 3pc für die schon genannte Mediathek der Stasi-Unterlagenbehörde ein neuartiges und modernes Online-Konzept. Die Behörde des Bundesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU) – wie sie korrekt bezeichnet wird – bewahrt in ihren Archiven die 1990 sichergestellten Unterlagen des Ministeriums für Staatssicherheit (MfS) der DDR auf. Mit mehr als 111 Kilometern Aktenmaterial, über 1,7 Millionen Fotos, zahlreichen Videos sowie Tonbändern aus den Abhörzentralen der Stasi handelt es sich um einen der größten Archivbestände in Deutschland [4].

Die grundlegende Idee des Online-Konzepts war es, den Nutzern verschiedene Zugangsmöglichkeiten für ein Archivmaterial zu liefern, das zunächst wenig ansprechend wirkte (Aktenmaterial) und sich zumindest für den Laien wenig verständlich las (Behördensprache). Unter Verwendung semantischer Technologien des Projektpartners Retresco [5] konnten die Archivalien für eine semantische Suche aufbereitet werden, um insbesondere die gezielte und fachkundige Suche im verfügbaren Aktenmaterial unterstützen zu können. Gleichzeitig konnten nun auf Basis der maschinell lesbaren Daten, die angereicherten Archivalien nach semantischen Bezügen geordnet werden, die einen eher emotionalen und explorativen Zugang ermöglichten.

Dazu wurden zum einen Arichivalien zu Themensammlungen mit redaktionellen Erläuterungen zusammengefasst, zum anderen wurde Aktienmaterial mit redaktionellen Texten zu Geschichten aufbereitet, um aufzuzeigen, welche persönlichen Schicksale in den Akten dokumentiert sind. Diese zunächst linear gestalteten Erzählstrecken (Storytelling) wurden nach dem Launch der Online-Mediathek um weitere explorative Navigationsmöglichkeiten zu semantisch ähnlichen Objekten erweitert (Semantic Storytelling) [6]. Was folgte waren Auszeichnungen wie der Designpreis in Silber im Wettbewerb „Gute Gestaltung“ des Deutschen Designer Clubs e. V. (2015) und der begehrte If Design Award 2016 in der Disziplin „3.0 Communication“ [7].

Vor diesem Hintergrund hat sich die Digitalisierung als ein große Chance erwiesen, weil sie ein bisher verborgenes Archiv zum ‚Sprechen‘ brachte. Bedenkt man allerdings, dass bis zum heutigen Zeitpunkt nur ein geringer Anteil des vorhandenen Aktenmaterials online verfügbar ist – zum jetzigen Zeitpunkt stehen online 2.500 Dokumentenseiten, 250 Einzelbilder, sechs Stunden Tonaufzeichnungen und 15 Stunden Filme online zur Verfügung – dann wird deutlich, dass die Digitalisierung eine noch zu bewältigende Herausforderung ist und bleibt.

2.2 DIGITALE BRIEFEDITIONEN

Ein weiteres Anwendungsgebiet, mit dem sich 3pc schon seit Jahren beschäftigt, ist im weitesten Sinne das Aufgabenfeld der Erschließung und Veröffentlichung digitaler Nachlass- und Autographensammlungen. Exemplarisch dafür steht das Projekt EMA, das digitale Erich Mendelsohn Archiv [8]. Der Architekt Erich Mendelsohn (1887-1953) gilt als einer der wichtigsten Wegbereiter und Vertreter der modernen Architektur. Sein reicher künstlerischer Nachlass verteilt sich auf die Archive der Kunstbibliothek – Staatliche Museen zu Berlin und des Getty Research Institute, Los Angeles. Ziel des Projekts war der Aufbau eines digitalen Archivs, um online einen integrierten Zugang zu den beiden räumlich getrennt verwahrten Beständen zu ermöglichen.

Im Mittelpunkt des Archivs steht der jahrzehntelange Briefwechsel zwischen Erich Mendelshohn und seiner Frau Luise. Dafür wurden im Projekt 1410 Erich- und 1328 Luise-Briefe digitalisiert, transkribiert und mit Anmerkungen versehen. Als Editionswerkzeug brachte 3pc in diesem Projekt den Refine!Editor zum Einsatz. Der Refine!Editor ist ein webgestütztes Werkzeug zur kollaborativen

Transkription, Indexierung und Online-Präsentation von Archivbeständen, das speziell für die Bedürfnisse im Bereich der wissenschaftlichen Erschließung von Handschriften entwickelt worden ist (Backend-Lösung). Er wurde von 3pc in Zusammenarbeit mit der Humboldt-Universität und der Staatsbibliothek zu Berlin entwickelt, um disloziertes und kollaboratives Transkribieren zu ermöglichen.

Mithilfe dieser webbasierten Editionssoftware ist es gelungen, eine digitale Edition zu erstellen, die gedruckten Verzeichnissen in vielerlei Hinsicht überlegen ist. Sie zielt nicht nur auf die Verfügbarmachung materieller Werke, sondern nutzt vielmehr im globalen Netz verfügbare elektronische Ressourcen, um die Texte insgesamt anzureichen. Zu diesen Anreicherungen zählen kritische Annotationen, eine umfassende Indexierung sowie die Unterlegung zahlreicher Begriffe und Werkbeschreibungen mit weiterführenden Links. So kamen Linked Data Technologien des Semantic Web zum Einsatz, um die digitale Edition mit dem „kartierten Universum geographischer Informationssysteme“ (Geonames) zu verbinden. Es wurden Normdaten der Deutschen Nationalbibliothek (DNB) und der Library of Congress (LOC) verwendet, um interoperable Referenzen zu Personennamen, Werkbezeichnungen oder Ereignissen aufzubauen. Und nicht zuletzt konnten Verlinkungen zu anderen Medienquellen wie SMB-Digital oder die Deutsche Digitale Bibliothek (DDB) eingepflegt werden, die dem Textmaterial eine höher Anschaulichkeit verliehen [9].

Auch dieses Beispiel macht deutlich, welche Chancen die Digitalisierung für innovative Formen der Verfügbarmachung und Publikation bietet. Es hat sich aber auch gezeigt, dass diese Form der Aufbereitung auch verbesserte Werkzeuge und ganze neue Redaktionsworkflows braucht, um den schier unfassbaren Umfang an digitalen Archivalien, den es in Zukunft noch zu erschließen gilt, bewerkstelligen zu können. Entsprechend konsequent wurde bei 3pc die Idee des Semantic Storytelling um das Ziel der Entwicklung eines smarten Autorensystems ergänzt und weitergedacht.

3. EIN SMARTES AUTORENSYSTEM

Worin besteht nun die Notwendigkeit für ein smartes Autorensystem und was sollte ein solches System leisten? Die Antworten leiten sich ab von den Erfahrungen wie sie in den beschriebenen Projekten gemacht worden sind. Sie lassen sich im Wesentlichen auf drei anwendungsbezogene Aspekte reduzieren:

Zeitgewinn
Die systematische Aufbereitung und Verfügbarmachung digitaler Archivalien ist ein aufwändiger Prozess, der nicht ohne manuelle Arbeiten und entsprechende Fachkenntnis zu bewerkstelligen ist. Ein smartes Autorensystem sollte Wissensarbeiter daher hauptsächlich bei Routine-Aufgaben wie beispielsweise der Verschlagwortung und Verlinkung unterstützen (semi-automatische Verfahren), um mehr Zeit für die kritische Editionstätigkeit zu gewinnen und vorhandenes Archivmaterial mit einem vertretbaren Kostenaufwand aufbereiten zu können.

Benutzerfreundlichkeit
Systeme zur Veröffentlichung von Archivalien im Netz basieren in der Regel auf webbasierten Content-Management-Systemen (CMS) wie beispielsweise Typo3. Sie sind meist nicht besonders benutzerfreundlich, was in jüngster Zeit – analog zum Begriff der User Experience (UX) – im Bereich des kommerziell motivierten Content Managements unter dem Begriff der Author Experience (AX) diskutiert wird [10]. Darüber hinaus müssen diese Systeme aufwändig an die jeweilige Aufgabenstellung angepasst werden, was die Kosten erhöht und die Bedienbarkeit des Systems eher erschwert denn verbessert. Ein smartes System sollte die Autoren entsprechend des ganzen Workflows von der Archivaufbereitung bis zu Publikation optimal unterstützen.

Storytelling-Formate
Wie sich gezeigt hat, sind Storytelling-Formate hoch geeignet, um digitale Kulturgüter insbesondere für ein breit aufgestelltes Publikum verfügbar zu machen. Gängige Redaktionssysteme verfügen jedoch nicht über entsprechende User Interfaces (UI), die Autoren sowohl bei der Recherche, als auch bei der Entwicklung einer Geschichte unterstützen. Das gewünschte Autorensystem sollte demnach über ein intuitives UI verfügen, das Funktionalitäten für beide Aufgabenstellungen unter einer Bedienoberfläche verfügbar macht.

3.1 REDAKTIONSWORKFLOW

Der Workflow zum Aufbau eines ansprechenden Online-Archivs bzw. einer Online-Mediathek lässt sich grob in folgende Arbeitsschritte einteilen:

1. Scannen
2. Transkribieren / OCR (Optical Character Recognition)
3. Katalogisieren (TEI-Standard)
4. Kommentieren (Annotieren, Verlinken etc.)
5. Online Publizieren (Index / Suche / Storytelling / Themensammlungen etc.)

Das von 3pc anvisierte smarte Autorensystem soll bis auf den Prozess des Scannes in Zukunft alle genannten Arbeitsschritte auf unterschiedliche Art und Weise unterstützen. Die auf Basis sogenannter Wirerframes bisher im Projekt entwickelte prototypische Benutzeroberfläche für das Autorensystem (vgl. Abb. 1) unterscheidet daher zwischen den Modulen Datenverwaltung (Archivaufbereitung) und Redaktionstool (Storytelling).

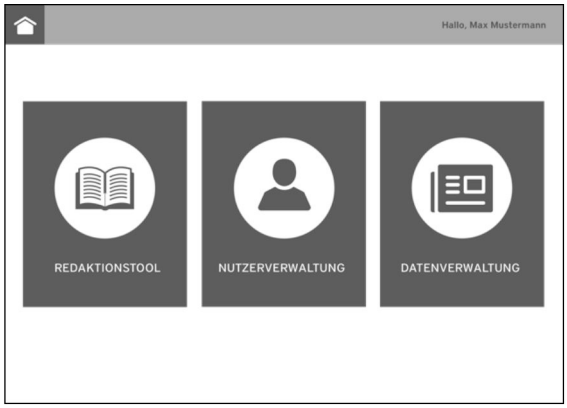

Abb. 1: *Startscreen des smarten Autorensystems (prototypische Benutzeroberfläche als Wireframe)*

Der Funktionsbereich des Moduls der Datenverwaltung soll künftig die Arbeitsschritte zwei bis vier unterstützen, das Redaktionstool insbesondere den Arbeitsschritt des Publizierens, wobei hier zu bemerken ist, dass das Kommentieren und Annotieren sich im Hinblick auf die angereicherten Digitalisate ebenfalls mit dem Arbeitsschritt des Publizierens überschneidet.

3.2 ARCHIVAUFBEREITUNG

Betrachtet man die Arbeitsschritte zwei bis vier, die über den Funktionsbereich der Datenverwaltung verfügbar sein sollen, wird schnell deutlich, an welchen Stellen ein smartes System den Wissensarbeiter effektiv unterstützen kann.

Im Forschungsprojekt DKT hat 3pc den Refine!Editor am Beispiel des digitalen Archivbestands des EMA-Projekts um prototypische smarte Funktionalitäten erweitert. Von Haus aus unterstützt das System bereits den Import von Digitalisaten und Datenquellen im TEI-XML-Format. Entsprechend der Digitalisierungsrichtlinie der DFG aufbereitete Digitalisate können so ohne großen Mehraufwand einfach in das System importiert werden. Darüber hinaus können über die Benutzeroberfläche für Editoren die eingescannten Digitalistate angezeigt und deren Inhalte über einen Texteditor transkribiert werden (vgl. Abb. 2).

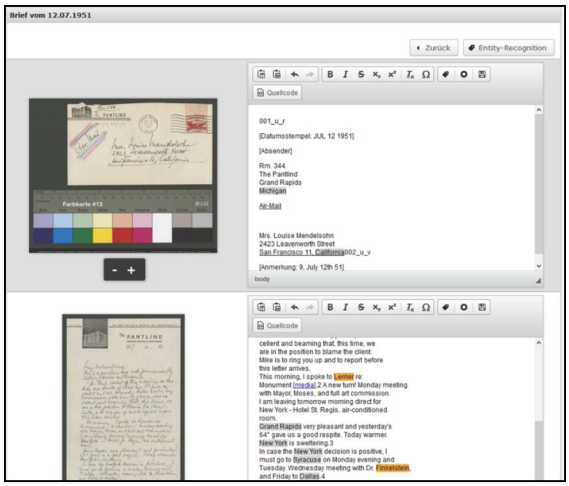

Abb. 2: *Erweiterte Benutzeroberfläche des Refine!Editors zur Transkription und semi-automatischen Annotation und Verlinkung (prototypische Implementierung)*

Die eigentliche Innovation des Systems betrifft jedoch die aufwändige manuelle Annotation von Begriffen wie Personennamen, Orte, Ereignisse usw. Hier setzt das System auf die Technologien zur sogenannten Named Entitiy Recognition (NER) des Forschungspartners DFKI. Per Mausklick können Editoren die automatische Erkennung und Verlinkung von Begriffen auf dem transkribierten Textmaterial anstoßen und anschließend die Annotationen auf ihre Qualität hin prüfen. Vorhandene Fehler können ggf. direkt editiert werden

(semi-automatisches Verfahren). Gleichzeitig werden die so annotierten Entitäten in einem maschinenlesbaren Format gespeichert (RDF/OWL) – die Basis für semantische Suchfunktionalitäten, automatisches Indexieren und maschinell gestütztes Storytelling (s. u.).

Die bisherigen Entwicklungsarbeiten am Prototypen haben gezeigt, dass diese Form der semi-automatischen Aufbereitung große Potenziale birgt, um einen spürbaren Zeitgewinn bei dieser Art der Tätigkeit zu erzielen. Darüber hinaus wirken sie sich umittelbar auf die anderen aufwändigen Aufgaben zur Indexierung und Verlinkung auf andere Online-Ressourcen aus. Gleichzeitig ist es möglich, bei ausreichend hoher Erkennungsqualität einen kompletten Archivbestand in einem Arbeitsschritt automatisch zu annotieren (Skalierbarkeit) und anschließend sorgfältig kritisch zu kommentieren.

Vor dem Hintergrund der bisherigen Ergebnisse der Forschungs- und Entwicklungsarbeiten konzentriert sich 3pc im nächsten Schritt nun gemeinsam mit dem Forschungspartner DFKI auf die Verbesserung des NER-Verfahrens, um die Erkennungsqualität durch Training auf den Daten, Feedback durch die Wissenarbeiter und die Einbeziehung weiterer Wissensquellen für den praktischen Einsatz optimieren zu können.

3.3 STORYTELLING

Es wurde bereits erwähnt, dass ein innovatives Redaktionssystem für Storytelling die Funktionalitäten Recherche und Erstellen von Geschichten (Schreiben, Bildredaktion etc.) in einem Autorentool unter einer intutiv bedienbaren Oberfläche vereinigen sollte. Die grundlegende Forschungsfrage lautete: Welche Vorteile bietet die semantische Aufbereitung der Daten in maschinenlesbarer Form für das Storytelling?

Antworten darauf wurden zunächst in Form von prototypischen Benutzeroberflächen (Wireframes) zur Unterstützung der Recherche in einem umfangreichen Archiv gesucht. Die ersten Entwürfe haben gezeigt, dass eine grafische Aufbereitung stichwortbezogener Recherchen nach biographischen, chronologischen und geographischen Bezügen eine vielversprechende Funktionalität ist. Sie hilft den Autoren dabei, Zusammenhänge in den Materialien zu erkennen, die sich auf den ersten Blick nicht sofort erschließen, unterstützt aber auch optimal bei der Expertensuche (vgl. dazu die Abb. 3 bis 5).

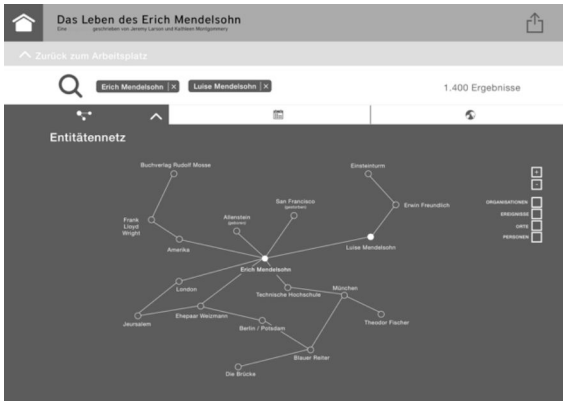

Abb. 3: Recherche nach biographischen Bezügen (Redaktionstool)

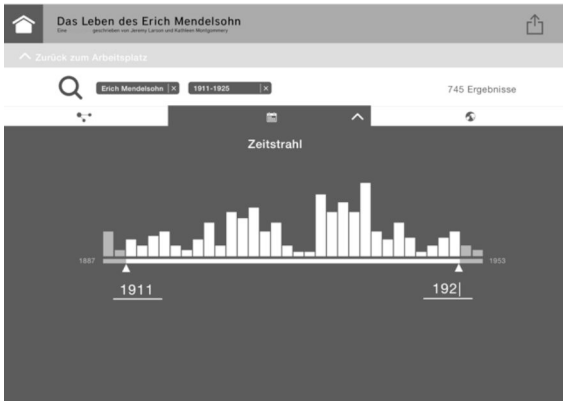

Abb. 4: Recherche nach chronologischen Bezügen (Redaktionstool)

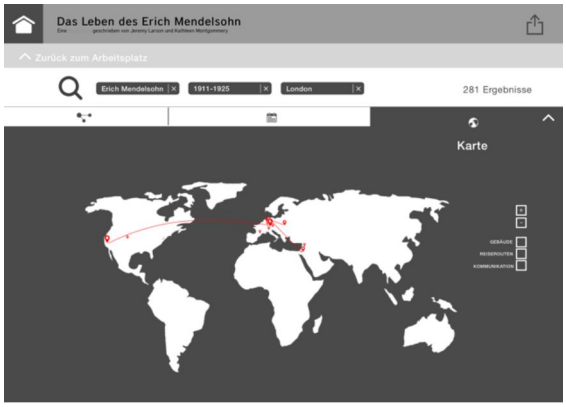

Abb. 5: Recherche nach geographischen Bezügen (Redaktionstool)

Während die geplanten Funktionalitäten zur Recherche im Hinblick auf die Realisierbarkeit schon recht weit durchdacht sind, haben die bisher entwickelten Benutzerobflächen für die Erstellung von Geschichten noch starken Studiencharakter. Ihr Kernanliegen ist es, auf Basis von thematischen Storytemplates (z. B. Kurzbiographie, Freundschaften, Reisen etc.) den Erstellungsprozess einer Geschichte

(Recherche, Schreiben, Bildredaktion etc.) möglichst intuitiv zu unterstützen. Dafür wurde eine Oberfläche entwickelt, die sich am Prinzip des WYSIWYG (what you see is what you get) orientiert. Sie soll es Autoren ermöglichen, recherchierte und in einer Merkliste zusammengestellte Archivalien zu einer multidimensionalen Storyline zusammenzustellen, die um eigene Texte und Überschriften ergänzt werden kann.

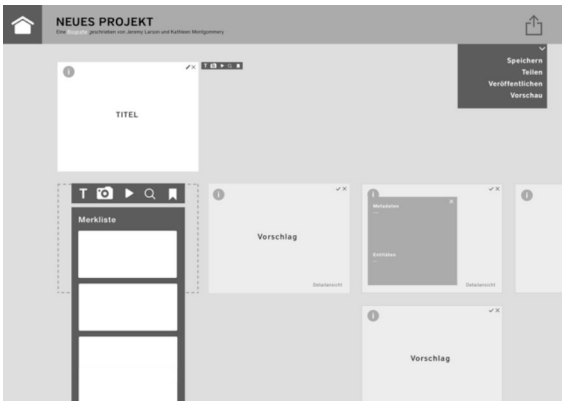

Abb. 6: *Oberfläche zur Erstellung von Storyelementen auf Basis recherchierter Archivalien auf der Merkliste (Redaktionstool)*

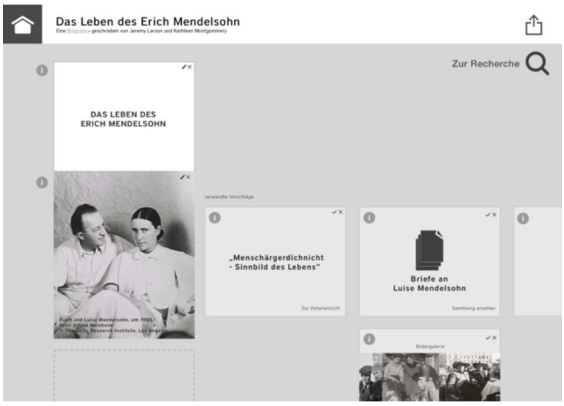

Abb. 7: *Multidimensionale Storyline (Redaktionstool)*

Die Gestaltung dieser Oberflächen ist nicht trivial und bedarf noch weiterer Entwicklungsschritte. Ziel ist es, am Ende des Entwicklungsprozesses den Autoren ein System zur verfügung stellen zu können, das ohne großen Schulungsaufwand nutzbar ist und den kreativen Prozess des Storytellings während der Erstellungsphase durch sinnvolle Empfehlungen thematisch relevanter Archivalien zu unterstützen.

4. AUSBLICK

Wie sieht der Redaktionsworkflow der Zukunft nach den bisher im Forschungsprojekt gemachten Erfahrungen nun aus? Eine definitive Antwort dazu kann es aus heutiger Sicht sicherlich nicht geben. Vor dem Hintergrund der derzeitigen allgemeinen Euphorie durch die in der Wissenschaft auf dem Gebiet der Künstlichen Intelligenz (KI) erzielten Fortschritte, werden semi-automatische Verfahren zur Archivaufbereitung (NER) und maschinelle Unterstützung beim Storytelling (Recherche / Empfehlungen) in den kommenden fünf Jahren sicherlich zum Standard werden.

Die Frage ist, ob die Potenziale der vorhandenen Sprach- und Wissenstechnologien durch ein sinnvolles Zusammenspiel mit den Wissensarbeitern und Kreativen auch voll entfaltet werden können? Ein entscheidendes Kriterium wird deren intuitive und sinnvoll erscheinende Benutzbarkeitkeit sein. Denn nichts ist weniger hilfreich als ein System, das seine Autoren mit einer schlechten „Experience" zurücklässt. Und eine Zeit, in der die Maschinen die Geschichten schreiben, erscheint zumindest im Bereich der Kulturarchive realistischerweise noch in weiter Ferne zu sein.

Anmerkung des Autors: Wenn Sie Interesse daran haben, unsere Prototypen einmal zu testen und Ihr Feedback zu geben, zögern Sie nicht, mich unter der angegeben E-Mail-Adresse zu kontaktieren. Wir freuen uns auf Ihre Resonanz!

5. REFERENZEN

1. Stasi Mediathek | Mediathek der Stasi-Unterlagen-Behörde
   http://www.stasi-mediathek.de/

2. Verbundprojekt Digitale Kuratierungstechnologien
   http://digitale-kuratierung.de/

3. Wikipedia-Artikel „Semantic Web“
   https://de.wikipedia.org/wiki/Semantic_Web
4. Ebd.
   http://www.stasi-mediathek.de/ueber-diese-seite/

5. Retresco GmbH
   http://www.retresco.de/

6. Vgl. beispielsweise die Geschichte *"Unter Kontrolle halten" Die Stasi und der Super-GAU von Tschernobyl*
   http://www.stasi-mediathek.de/geschichten/unter-kontrolle-halten/sheet/0-0/type/cover/

7. *If Design Award 2016: Wieder eine Auszeichnung für die "Stasi-Mediathek"*
   (News vom 02.02.2016)
   http://3pc.de/presse/item.html?id=305

8. EMA - Erich Mendelsohn Archiv
   Der Briefwechsel von Erich und Luise Mendelsohn 1910-1953
   http://ema.smb.museum/

9. Vgl. dazu Bienert, Andreas, *EMA - Erich Mendelsohn Archiv – Online editieren*,
   http://ema.smb.museum/de/projekt/online-editieren

10. Kraft, Boris, *Autoren dieser Welt vereint Euch! (Nieder mit schlechten Autorenumgebungen!)*
    http://www.contentmanager.de/cms/enterprise-cms/autoren-dieser-welt-vereint-euch-nieder-mit-schlechten-autorenumgebungen/