Heinrich C. Mayr, Martin Pinzger (Hrsg.): INFORMATIK 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 2113

new/s/leak - Anforderungsanalyse einer interaktiven Visualisierung für Data-Driven Journalism

Franziska Lehmann[1]

Abstract: Für Datenjournalisten[2] wird es immer schwieriger und zeitaufwändiger, die steigende Menge an Dokumenten zu analysieren und die für sie relevanten Informationen zu extrahieren. Mit Hilfe von Natural Language Processing können die in den Dokumenten vorkommenden Personen, Organisationen und Orte, sogenannte Entitäten, sowie deren Verbindungen identifiziert werden. Die Exploration der extrahierten Entitäten wird durch interaktive Visualisierungen unterstützt. Diese Idee wird bereits durch das Tool "Netzwerk des Tages" umgesetzt. Allerdings benötigt das Programm wesentliche Verbesserungen, um die journalistischen Bedürfnisse zu erfüllen. Auf dieser Basis erfolgte im Rahmen eines Projektes für das Tool "new/s/leak" eine Anforderungsanalyse mit Experten eines großen, deutschen Verlagshauses. Dieser Beitrag behandelt die in den Interviews identifizierten wesentlichen Anforderungen.

Keywords: Data-driven Journalism, Anforderungsanalyse, Informationsvisualisierung, Visual Analytics, Experteninterviews

1 Einleitung

"Now that we're deep into the information age, it's time for everyone to accept that the amount of information in our lives is only going to keep growing" (Mark Briggs [Br09])

Im Zeitalter der Globalisierung und Digitalisierung entstehen immer größer werdende Datenbestände. Das Ziel von investigativen Journalisten ist die Identifizierung bisher unbekannter, bedeutender Muster und Beziehungen innerhalb dieser Datensammlungen sowie deren Veröffentlichung [Lu07]. Die enormen Datensammlungen stellen hierbei eine große Herausforderung dar. Denn neben strukturierten Daten beinhalten die Dokumentensammlungen normalerweise vor allem unstrukturierte Textdokumente, wie im Falle von Wikileaks [Gy14] oder der Panama Papers. Die Exploration der Daten sowie die Identifizierung wichtiger Dokumente ist schwierig und zeitaufwändig. Dies ist insbesondere der Fall, wenn Journalisten im Voraus nicht wissen, wo sie ihre Recherche beginnen sollen, was in den Dokumenten wichtig sein könnte oder wie Ereignisse im Zusammenhang stehen [Br14, Gö13]. Zudem müssen die Texte aufgrund des Wettbewerbs zwischen Medienhäusern unter Zeitdruck analysiert werden [UK15]. Somit steigt der Bedarf an explorativen Datenanalysetools zur Unterstützung der journalistischen Recherche.

[1] Technische Universität Darmstadt, f.lehmann@stud.tu-darmstadt.de
[2] In dieser Arbeit wurde immer die männliche Form von Personen benutzt, damit der Lesefluss nicht gestört wird. Die Personen umfassen dementsprechend auch weibliche Personen.

Die Erstellung eines journalistischen Artikels basiert auf der Beantwortung der sog. fünf W-Fragen der journalistischen Recherche: 'Wer?', 'Was?', 'Wo?', 'Wann?' und 'Warum?' [Zh13]. Das Untersuchen und Entdecken der in den Dokumenten enthaltenen Themen, Inhalte sowie das Verständnis von Verbindungen und Beziehungen zwischen den darin genannten Personen, Organisationen und Orten (Entitäten) ist somit ein wichtiger Bestandteil des investigativen Prozesses [Gö13]. Allerdings kann dies aufgrund der meist enormen Menge an Dokumenten ohne computergestützte Analyse zeitaufwändig sein [Br14, Gö13].

Zur Lösung dieser Herausforderung können zum einen auf Basis einer **computergestützten Analyse** mit Hilfe von Data-Mining Methoden Modelle zur Charakterisierung der Daten berechnet werden, zum anderen können Journalisten mittels **Informationsvisualisierung** direkt mit der visuellen Schnittstelle interagieren und den Datensatz analysieren [Gö13, Su13]. Die Kombination und Interaktion visueller und automatischer Analyse ist Gegenstand von Visual Analytics, einem interdisziplinären Ansatz, der die Vorteile aus den beiden Forschungsgebieten verbindet [Su13]. Neben den unzähligen Tools zur Analyse von strukturierten Daten [Jä15], existieren nur wenige Tools für die Analyse von unstrukturierten Daten, die Entitätenextraktion mit visuellen Interaktionsmöglichkeiten bieten [Gö13, Su16]. Diese weisen allerdings verschiedene Nachteile auf (siehe hierzu Abschnitt 2.2), repräsentieren bzw. beantworten die fünf W-Fragen unzureichend und funktionieren bisher nur mit englischsprachigen Texten.

An der Schnittstelle von visueller und automatischer Analyse setzt das Projekt "Data Extraction and Interactive Visualization of unexplored Textual Datasets for Investigative Data-Driven Journalistic (DIVID-DJ)" an. Dies ist ein interdisziplinäres Projekt des Fachgebiets Graphisch-Interaktive Systeme mit dem Fachgebiet Sprachtechnologie der TU Darmstadt und wird in Zusammenarbeit mit einem großen, deutschen Verlagshaus durchgeführt. Das Ziel des genannten Projektes ist die Entwicklung eines Programms zur Unterstützung von Datenjournalisten. Das Tool "network of searchabe leaks" (new/s/leak) soll dem Journalisten als Hilfsmittel dienen, um in zeitkritischen Situationen tiefe Einblicke in neu gewonnene Textdokumente zu erhalten.

Die Basis des Tools bildet das System "Netzwerk des Tages" [Fa14, KLB14], das aus einem vorigen Projekt der genannten Fachgebiete resultierte. Mit diesem Tool ist schon jetzt eine Exploration von Entitätenbeziehungen möglich. Ein beispielhafter Netzwerkausschnitt ist in Abb. 1 dargestellt.

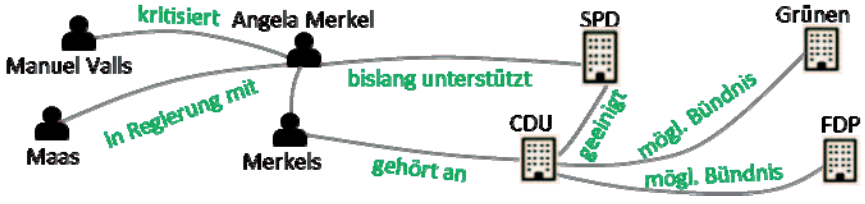

Abb. 1: Ausschnitt eines Entitätennetzwerks aus dem "Netzwerk des Tages"

Bereits der Projektantrag enthielt Darstellungselemente, sogenannte Views, die für das Tool new/s/leak geplant sind, wie bspw. einer Zeitleiste, einer aggregierten Sicht der Entitäten sowie einer Dokumentenansicht. Damit das Programm entsprechend der journalistischen Bedürfnisse entwickelt werden kann, müssen zum einen die geplanten Views geprüft und konkretisiert, sowie weitere Anforderungen innerhalb einer Anforderungsanalyse identifiziert werden. Der vorliegende Beitrag greift daher die weiter oben aufgezeigte Forschungslücke auf und präsentiert die Analyse und die identifizierten Anforderungen bzgl. der interaktiven Visualisierung des Tools.

Der Beitrag ist wie folgt gegliedert: In Abschnitt 2 werden die relevanten journalistischen Begrifflichkeiten definiert sowie die in der Literatur identifizierten Visual Analytics Tools in Bezug auf die vorliegende Problemstellung vorgestellt. Abschnitt 3 enthält die Beschreibung der Vorgehensweise der Anforderungsanalyse. In Abschnitt 4 erfolgt die Beschreibung der wichtigsten Anforderungen, extrahiert aus den Interviewergebnissen. Der Beitrag endet mit einem Fazit und Ausblick in Abschnitt 5.

2 Journalistischer Kontext und verfügbare Tools

Dieser Abschnitt behandelt die relevante journalistische Terminologie sowie relevante Tools aus dem Bereich Visual Analytics.

2.1 Datenjournalismus

Datenjournalismus wird als die Kombination eines Recherche-Ansatzes und einer Veröffentlichungsform herausgestellt, indem maschinenlesbare Datensätze durch die Verwendung von Software miteinander verschränkt und analysiert werden [Ma10]. Die Analyse der Daten erfolgt anhand des experimentellen Einsatzes von Algorithmen, Daten und sozialwissenschaftlichen Methoden [Au15]. "[Damit] wird ein schlüssiger und vorher nicht ersichtlicher informativer Mehrwert gewonnen. Diese Information wird in statischen oder interaktiven Visualisierungen angeboten und mit Erläuterungen zum Kontext, Angaben zur Datenquelle (bestenfalls wird der Datensatz mit veröffentlicht) versehen." [Ma10] Hierbei deckt Datenjournalismus den journalistischen Arbeitsprozess, von der Datensammlung, Datenanalyse und Filterung, Visualisierung sowie Berichterstattung, ab und ermöglicht dem Leser den Einstieg in große Datensätze [Au15].

2.2 Verfügbare Tools

In der Literatur zu Visual Analytics wird eine Vielzahl an Tools beschrieben, die sich auf einen bestimmten Aspekt der in diesem Beitrag behandelten Problemstellung fokussieren.

Die Tools **PaperLens** [Le05] und **NetLens** [Ka06] unterstützen die Filterung der Dokumente nach Metadaten in Form von Balkendiagrammen, Listen, Grafiken und textbasierten Visualisierungen zur Darstellung von Autor, Thema von Dokumenten sowie Zitationsdaten. Mit Hilfe der Identifikation der Themenbereiche erfolgt ebenso eine Gruppierung der Dokumente. Aufgrund der fehlenden Entitätenerkennung sowie deren Verbindungen müssten Journalisten immer noch einen Großteil der Dokumente lesen.

Die folgende Gruppe von Tools fokussiert sich insbesondere auf die Darstellung inhaltlicher Themengebiete der vorliegenden Dokumentsammlung. Während **Overview** [Br14] lediglich den Inhalt der Dokumente als eine Menge von Schlüsselworten visualisiert, stellen **ThemeRiver** [Ha00] sowie **Parallel Topics** [Do11] zusätzlich die zeitliche Veränderung von Themen, extrahiert aus der Dokumentensammlung, dar. Diese Tools bieten allerdings keine Entitätenerkennung an, weshalb sie angesichts des Mangels an intuitiver und präziser Visualisierung zur Untersuchung der vorliegenden Problemstellung ungeeignet scheinen.

Die Erkennung und Visualisierung von Entitäten wird u.a. durch die nachfolgenden ausgewählten Tools unterstützt. **Open Calais** [Ga13] unterstützt die Zusammenfassungen (Aggregationen) von Entitäten und markiert sie im Text. **BiSet** [Su16] und **Jigsaw** [Gö13] visualisieren die extrahierten Entitäten in einer Listenansicht und stellen Beziehungen zwischen Entitäten mit Verbindungslinien dar. Hingegen visualisiert das **Netzwerk des Tages** [Fa14, KLB14] die extrahierten Entitäten als Node-Link-Diagramm. Diese Tools bieten entweder keine Zeitleiste als Visualisierungsmöglichkeit an oder erfassen nur Zeitpunkte, benötigen einen zu hohen Schulungsaufwand bzgl. der Nutzung der Tools und/oder beinhalten keine bzw. geringe Filtermöglichkeiten. In der nachfolgenden Tabelle sind die möglichen Nachteile der Tools aufgeführt.

Tool	Mögliche Nachteile
PaperLens/NetLens	keine Entitätenerkennung, großer Leseaufwand
Overview	keine Entitätenerkennung, fehlende Zeitleiste
ThemeRiver/Parallel Topics	keine Entitätenerkennung sowie Verbindung derer
Open Calais	fehlende Zeitleiste, Schulungsaufwand
BiSet	fehlende Zeitleiste
Jigsaw	fehlende Filtermöglichkeiten
Netzwerk des Tages	Anzeige Netzwerk nur für einen Zeitpunkt

Tab. 1: Zusammenfassung möglicher Nachteile der Tools

3 Vorgehen der Anforderungsanalyse

Um eine möglichst vollständige Liste der Anforderungen an das Tool "new/s/leak" zu

erhalten, wird die Designstudienmethodik (engl. Design Study Methodology) als Grundgerüst für ein strukturelles Vorgehen zur Beantwortung der Forschungsfrage genutzt. Eine Designstudie ist ein Projekt, in dem ein Forscher ein reales Problem, dem ein Domänenexperte gegenübersteht, analysiert, zur Problemlösung ein Visualisierungssystem entwirft, diesen Entwurf validiert und die gewonnenen Erkenntnisse zur Verfeinerung der Designrichtlinien reflektiert und wird durch einen 9-stufen Rahmen praktisch angeleitet [SMM12]. Dieser Beitrag konzentriert sich auf die Stufe 'Erkenntnisphase'.

Die '**Erkenntnisphase'** schließt die Anforderungsanalyse ein [SMM12]. Das wichtigste Hilfsmittel von Anforderungsanalysten zur Anforderungsermittlung sind Interviews mit Nutzern und Stakeholdern. Das Ziel des Gesprächs ist die Erkennung der spezifischen Bedürfnisse von Nutzern an das Tool. Dies umfasst ebenso zu verstehen, welche Bedingungen erfüllt sein müssen, um das Tool für die tägliche Arbeit zu nutzen [SG06].

Die Analyse der Anforderungen erfolgt anhand nichtstandardisierter Leitfadeninterviews mit Experten [MN09]. Leitfadeninterviews enthalten vorgegebene Fragen, deren Reihenfolge und Formulierung unverbindlich gestaltet ist. Hierdurch können mit einem möglichst natürlichen Gesprächsverlauf gezielt Themen angesprochen und neue Gesichtspunkte identifiziert werden [GL10]. Bei Experteninterviews steht zudem nicht der Befragte als Person, sondern seine Erfahrungen und Interpretationen bzgl. des entsprechenden Forschungsthemas im Vordergrund [BG09]. Demnach verfügen Experten über Spezialwissen, das nicht Jedem in dem interessierenden Handlungsfeld zugänglich ist [MN09].

Als Grundlage der Anforderungsanalyse existiert bereits ein geplanter, grober Rahmen, anhand dessen die Anforderungen abgeleitet werden können. Der Rahmen beinhaltet verschiedene Sichten (Views) der Dokumente: eine Übersicht, eine Zeitleiste, ein Netzwerk sowie eine Dokumentenansicht. Durch die Interviews werden die Views überprüft, konkretisiert und entsprechende Anforderungen abgeleitet. Um die Anforderungsanalyse zu erleichtern und Gedanken zu konkretisieren wird zudem die Methode Sketching der Five Design Sheet (FDS) Methodik verwendet [RHR16]. Die FdS Methodik entspricht einer Komplettlösung, die divergentes Denken fördert, auf dem Skizzieren basiert und den Fokus auf das Ziel des zu entwerfenden Tools, seine Operationen und Interaktionen lenkt.

Zur Erstellung der Fragen des Interviewleitfadens wurden Informationen aus dem Projektantrag sowie Informationen über artverwandte Tools aus der Literatur genutzt. Der Leitfaden enthält zwei größere Abschnitte. Der erste Teil beinhaltet Fragen zur derzeitigen journalistischen Arbeitsweise. Der zweite Abschnitt enthält Fragen, welche direkt das Tool betreffen. Der grundsätzliche Aufbau des zweiten Teils orientiert sich an den Darstellungselementen (Views), die im Projektantrag genannt wurden: Übersicht, Netzwerk, Suche, Zeitinformationen/Zeitleiste, Dokumentenansicht, Annotationen. Hinzu kamen die Kategorien Geo-Visualisierung, Vergleiche sowie Anzeigedauer. Die Interviews fanden mit fünf Datenjournalisten in drei separaten Interviews statt.

Die Datenanalyse und Auswertung orientiert sich an dem von Gläser und Laudel (2010) vorgestellten Verfahren, der qualitativen Inhaltsanalyse. Hierbei erfolgt eine Extraktion von Informationen aus den Interviews hinsichtlich eines systematischen Schemas. Dadurch wird eine frühzeitige Separation vom Ausgangstext erreicht, die nur Informationen enthält, welche für die Anforderungsanalyse von Interesse sind [GL10].

4 Ergebnisse: relevante Anforderungen

Im folgenden Abschnitt erfolgt die Vorstellung der relevanten Anforderungen, die sich aus der Anforderungsanalyse ergeben haben. In Tabelle 2 sind die wesentlichen Anforderungen aufgeführt. Die letzte Spalte der Tabelle enthält die Nennung der Anforderungen innerhalb den Interviews (eine Nennung pro Interviewteilnehmer möglich). Die erste Zahl beschreibt die Anzahl der Nennungen und die zweite Zahl, wie viele Interviewteilnehmer, bedingt durch den Leitfaden, die Anforderung genannt haben. Die Beschreibung der Anforderungen findet nach der Tabelle statt.

ID	Anforderung	Kurzbeschreibung	Ø
A1	Suche	Funktion, um bestimmte Begriffe in den Dokumenten zu finden	5/5
A2	Tagging	Hinzufügen von Metainformationen an Dokumente	5/5
A3	Filterung	Filterung der Dokumente nach bspw. Personen, Schlüsselwörtern oder Orten	4/4
A4	Interessantheitseinstellung	Einstellbarkeit der angezeigten Entitäten nach der Häufigkeit	4/4
A5	Interaktion zwischen Views	Durch Veränderung von Informationen einer View verändern sich ebenso die angezeigten Informationen anderer Views	4/4
A6	Export Dokumente	Export von interessanten Dokumenten als bspw. Archivdatei	4/4
A7	Annotationen	Funktion, um im Netzwerk/in Dokumenten Kommentare und Markierungen zu machen	4/4
A8	Bearbeitungsoptionen der Netzwerkkomponenten	Manuelle Beseitigung von Fehlern im Netzwerk durch Zusammenfassung, Löschung, Bearbeitung und Hinzufügen von Entitäten	1/1
A9	Übersichtsansicht	Anzeige der Metadaten der Dokumente, häufigste Entitäten und Schlüsselworte	5/5
A10	Netzwerk	Anzeige von Entitäten, deren Verbindungslinien und -kontexte & Netzwerkkennzahlen	3/5
A11	Ausschaltbare Netzwerkelemente	Ausblendbare Informationen innerhalb des Netzwerkes	4/4
A12	Zeitleiste	Anzeige von Dokumentenhäufigkeit,	4/4

		Entitäten, Beziehungen, Schlüsselwörter über Zeit	
A13	Dokumentenanzeige	Direkter Zugriff auf Dokumente durch Anzeige von Überschrift plus Textausschnitt, Metadaten und Schlüsselworte der Dokumente	5/5
A14	Karte	Visualisierung von Orten auf der Karte sowie Karte als Filterungsmöglichkeit nach Orten	4/4
A15	Not getting lost	Gestaltung der interaktiven Visualisierung, dass Journalisten nicht den Überblick verlieren	3/3

Tab. 2: Übersicht der wichtigsten Anforderungen

Suche: Die Suche ist mitunter eine der wichtigsten Anforderungen an das Tool und wurde innerhalb der Interviews von jedem Teilnehmer angesprochen. Die Suche nach Namen, Ländern oder Schlüsselwörtern ist eine gängige Vorgehensweise bei der Recherche innerhalb einer großen Menge an Dokumenten, wie bspw. im Falle der Wikileaks. Die durch die Suche getätigte Filterung ermöglicht einen wertvollen Informationsgewinn, z.B. wie sich Häufigkeitsdiagramme von Entitäten oder die Zeitleiste ändern. Informationen, die zusätzlich angezeigt werden sollten, sind u.a. die Häufigkeit des Suchbegriffs (verbleibende Anzahl der Dokumente nach Filterung) oder wie oft der Suchbegriff in Verbindung mit einem Namen in den Dokumenten vorkommt.

Tagging: Auch der Bedarf einer Tagging-Funktion konnte aus den Interviews (in-)direkt abgeleitet werden. Mit Hilfe der Tagging-Funktion können Journalisten Metainformationen an Dokumente anhängen, um diese bspw. in unterschiedliche Themenbereiche einzusortieren. Weiterhin ist damit die Möglichkeit verbunden, Dokumente oder relevante Teile des Netzwerkes mit einem Kollegen zu teilen. Da diese Funktion auch genutzt werden soll, um interessante Dokumente mit anderen Kollegen teilen zu können, ist es von Vorteil Notizen an die Tagging-Gruppe anhängen zu können.

Filterung: Eine der wichtigsten Anforderungen, die aus allen Interviews hervorgeht, betrifft die Filterung der Dokumente nach unterschiedlichen Kriterien. Die Visualisierung wird als ein einfach zu bedienender Filter für die Suche nach berichtenswerten Dokumenten benötigt. Durch die Anwendung von Filtern kann die Menge an Dokumenten reduziert werden. Dabei ist es vor allem wesentlich, Wichtiges von Unwichtigem zu trennen und einen Überblick zu bekommen, was wirklich relevant ist. Im Verlauf der einzelnen Interviews konnten mehrere Kriterien identifiziert werden, nach denen eine Filterung der Dokumente möglich sein sollte. Nachfolgend ist eine Auswahl der identifizierten Kriterien aufgeführt: aus den Dokumenten extrahierten Personen, Organisationen und Orte (Entitäten); Schlüsselwörter, die besonders oft in den Dokumenten vorkommen; Zeitspanne/Zeitpunkt; Dateiformate; Dateigröße sowie eine durch den Journalisten vergebene Metainformation (Tag).

Interessantheitseinstellung: Was für Journalisten innerhalb einer Dokumentensammlung interessant ist, fällt je nach Dokumentensammlung und welche Fragen Journalisten beantwortet haben wollen, unterschiedlich aus. Dementsprechend

gibt es Fälle, in denen häufig auftretende Entitäten interessant sein können sowie Fälle, in denen selten auftretende Entitäten interessant sind. Um die unterschiedlichen Betrachtungsweisen zu unterstützen, ist die Einstellbarkeit der angezeigten Entitäten nach der Häufigkeit wichtig.

Interaktionen zwischen Views: Ebenso wurde die Interaktion zwischen Views in den Interviews mehrfach genannt. Die Interaktion zwischen den Views meint hierbei, dass bspw. durch das Setzen einer Filterung sich das Netzwerk ändert. Ein weiteres Beispiel ist, dass beim Anwählen eines Knoten innerhalb des Netzwerkes zusätzlich die Zeitleiste des Knotens angezeigt werden soll. Dementsprechend wird die Interaktion der Filter mit den Views und den darin enthaltenen Elementen benötigt. Dies umschließt u.a. die Zeitleiste, die angezeigten Häufigkeitsdiagramme, das Netzwerk, die Schlüsselwörter und weitere Informationen, die für die getätigte Filterung bestehen.

Export Dokumente: Bei der Zusammenarbeit von Journalisten mit anderen Fachkollegen, werden relevante Dokumente untereinander ausgetauscht. Aufgrund dieser Gegebenheit kam der Wunsch auf, Dokumente exportieren zu können. Daher müssen Dokumente im ersten Schritt auswählbar gemacht werden, um diese im zweiten Schritt bspw. als Archiv exportieren zu können.

Annotationen: Damit Journalisten Verknüpfungen zwischen Dokumenten herstellen, bestimmte Passagen im Dokument optisch hervorheben (highlighten) oder einzelne Textpassagen mit ausführlichen Kommentaren versehen zu können werden Annotationsmöglichkeiten benötigt. Hierbei sollen Annotationen nicht nur in Dokumenten, sondern auch im Netzwerk möglich sein.

Bearbeiten von Netzwerkkomponenten: Die automatische Extraktion von Entitäten inkl. deren Verbindung mit Hilfe von Natural Language Processing ist mit Unsicherheit behaftet und somit fehleranfällig. Somit können Entitäten im Dokument nicht erkannt werden, der Typ einer Entität wird falsch erkannt oder unterschiedlich formulierte Entitäten, die jedoch die gleiche Bedeutung besitzen, werden nicht als eine einzige Entität extrahiert. Zur Beseitigung dieser Fehler sind daher verschiedene Bearbeitungsfunktionen notwendig, wie die Zusammenfassung zweier Entitäten mit der gleichen Bedeutung, das Löschen von Entitäten, das Hinzufügen von Entitäten zum Netzwerk oder der Bearbeitung einer Entität (Beschriftung sowie Metadaten wie bspw. der Typ einer Entität).

Informationen zur Übersicht: Damit der Journalist einen Einblick bekommt, welche Themen in Dokumentensammlungen enthalten sind, welche Personen vorkommen und welche Zeitspanne die Dokumentensammlung beinhaltet, werden zunächst die wichtigsten Informationen aus den Dokumenten benötigt. Die Anzeige von Häufigkeitsdiagrammen von bspw. Personen, Organisationen, Orten, Schlüsselwörtern ist somit eine zentrale Anforderung an das Tool. Als Visualisierung der häufigsten Entitäten werden schlichte Balkendiagramme präferiert. Zudem stellte sich die Anzeige der Metadaten der Dokumente innerhalb aller Interviews als bedeutende Information über den Datensatz heraus. Die in Dokumenten vorkommende Metadaten sind u.a. die

Quelle, von der die Dokumente stammen; das Sicherheitslevel der Dokumente; der Dateiname des Dokuments; das Dateiformat der Dokumente; die Dateigrößen; wie viele Dokumente in der Dokumentensammlung enthalten sind sowie dem Autor der Dokumente.

Netzwerk: Das Netzwerk stellt eine zentrale View des Tools 'new/s/leak' dar. Darin sind die aus den Dokumenten extrahierten Entitäten aufgeführt. Neben den Entitäten wird die Anzeige der Verbindungslinien zwischen zwei Entitäten benötigt. Die Verbindungslinie signalisiert hierbei die Existenz von Beziehungen zwischen Entitäten, extrahiert aus den darunter liegenden Dokumenten. Allerdings ist neben der Verbindungslinie auch der Inhalt der Verbindungslinie für die Journalisten von Interesse und verleiht der Verbindungslinie mehr Aussagekraft. Durch die Darstellung der Entitäten und deren Verbindungen mittels eines Netzwerkes können zudem weitere Informationen berechnet und extrahiert werden. So wurden in den Interviews Kennzahlen genannt, die Fragen abbilden wie wer die meisten Verbindungen hat oder wo im Netzwerk dichte Gruppen existieren. Die Anzeige und Filterung nach bspw. denjenigen Entitäten, die die meisten Verbindungen zu anderen Entitäten aufweisen, sollte daher durch geeignete Kennzahlen umgesetzt werden.

Ausschaltbare Netzwerkelemente: Werden alle Elemente von Anfang an angezeigt, kann dies zur Überladung des Netzwerkes führen. Da die Journalisten unterschiedliche Informationen zu verschiedenen Zeitpunkten angezeigt bekommen möchten, ist es notwendig, Netzwerkelemente ausschalten zu können. Hierbei bilden die extrahierten Entitäten die Basis des Netzwerkes. Die weiteren Netzwerkinformationen wie die Verbindungslinien, der Inhalt einer Verbindungslinie oder die Gewichtung von Knoten sollen als hinzufügbare Elemente auswählbar sein.

Zeitleiste: Die Zeit ist eine der am häufigsten genannten Information innerhalb der Interviews. Mit Hilfe der Zeitleiste sehen Journalisten auf einen Blick den zeitlichen Umfang der Dokumente (definiert über die Erstellungsdaten des ältesten und jüngsten Dokuments der Sammlung) sowie die Verteilung von Dokumenten über die Zeit. Anhand der Zeitleiste lassen sich zu verschiedenen Zeitpunkten der Recherche unterschiedliche Informationen darstellen, die im Zusammenspiel mit dem Netzwerk und der Dokumentenansicht interessante Einblicke in die Daten geben können. Informationen, die entlang der Zeitachse für die Journalisten von Bedeutung sind, ist die Dokumentenhäufigkeit, das Vorkommen der Entitäten über die Zeit und deren Verbindungen sowie die Anzeige der häufigsten Schlüsselwörter über die Zeit.

Dokumentenanzeige: Eine weitere View, die Journalisten benötigen, ist die Dokumentenansicht. Hierdurch wird sichergestellt, dass die Journalisten immer direkten Zugriff auf die Dokumente der Dokumentensammlung haben und sehen können, welche Dokumente nach getätigter Filterung noch übrig bleiben. In der Dokumentenansicht werden Informationen über die Dokumente der Dokumentensammlung benötigt. Eine der wichtigsten Informationen, ist die Anzeige der Überschrift der Dokumente inklusive der ersten Zeilen. Mittels des Anrisses der Dokumente verschafft sich ein Journalist

einen Überblick über die Dokumente und muss nicht jedes einzelne durchlesen. Daneben sind die Metadaten der Dokumente ein wichtiger Bestandteil der Dokumentenübersicht. Die Metadaten können die Informationen Dateiname, Dateiformat, Dateigröße sowie Autor umfassen. Aber auch die Schlüsselworte eines Dokumentes wird für die Dokumentenanzeige benötigt, weil die Journalisten hierdurch schnell einen Überblick bekommen, welche Dokumente spannend und welche nicht spannend sein könnten.

Karte: Einer geographischen Karte wurde innerhalb der Interviews eine eher nachrangige Bedeutung beigemessen. Hierdurch könnte die Karte als "nice-to-have" Funktion gesehen werden. Aufgrund der wenigen durchgeführten Interviews und dem Wissen, dass außerhalb der Interviews nach einer Karte als Visualisierung explizit gefragt wurde, wird deren Visualisierung trotzdem als Anforderung aufgenommen. Einerseits können anhand der Karte Orte und Datenpunkte visualisiert werden, die innerhalb der Dokumente vorkommen. Andererseits wird die Option benötigt, um nach Ländern und Städten filtern zu können.

Not getting lost: Eine nicht-funktionale Anforderung betrifft den Anspruch, in der Visualisierung nicht den Überblick zu verlieren. Im Journalismus wird in der Recherche ständig der Blickwinkel gewechselt, von der Metaebene in die Mikroebene und wieder zurück. Daher ist es wichtig den Prozess des investigativen Journalismus so gut wie möglich zu unterstützen.

In Abb. 2 sind die meisten der vorgestellten Anforderungen in einem Designvorschlag des Tools umgesetzt. Die roten Kreise beinhalten die IDs der jeweiligen Anforderungen.

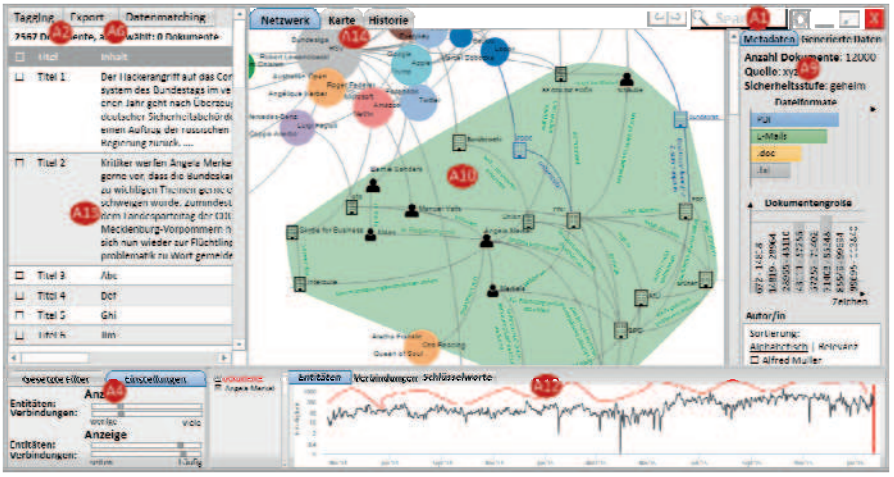

Abb. 2: Umsetzungsbeispiel der Anforderungen

5 Fazit

New/s/leak ist ein Visual Analytics Tool zur Unterstützung von Datenjournalisten bei der Exploration großer Datensammlungen. Der vorliegende Beitrag behandelt die relevanten Anforderungen bzgl. der interaktiven Visualisierung des Tools, identifiziert aus den durchgeführten Experteninterviews. Die Filterung der Dokumentensammlung anhand von Metadaten, Zeit, Entitäten und Schlüsselwörtern ist wichtig, um die Menge an Dokumenten in kürzester Zeit zu minimieren. Allerdings darf zwischen dem Journalist und der Dokumentensammlung keine Barriere vorhanden sein, sodass eine permanente Darstellung der Dokumentenanzeige benötigt wird. Da die Exploration von Datensammlungen im Team sattfindet, ist das einfache Kennzeichnen von relevanten Dokumenten durch Vergabe von Tags eine weitere der von Journalisten benötigten Funktionen. Aufgrund der automatisierten Entitätenerkennung können die Netzwerke Fehler enthalten. Daher ist es sinnvoll, verschiedene Bearbeitungsfunktionen des Netzwerks anzubieten, wie das Löschen von Entitäten. Zudem interessierten sich die interviewten Datenjournalisten außerdem für die Berechnung von Netzwerkkennzahlen. Dadurch lässt sich das Netzwerk bspw. nach denjenigen Entitäten filtern, die die meisten Verbindungen zu anderen Entitäten aufweisen.

Die nächsten Schritte des Projektes beinhalten die Implementierung der Anforderungen im bisher existierenden Tool sowie die Evaluierung des bisherigen Prototyps durch Nutzerstudien. Da sich die herausgearbeiteten Anforderungen insbesondere auf derzeitige Abläufe sowie bekannte Visualisierungen zurückführen lassen und zukunftsweisende, innovative Gestaltungselemente nicht ausreichend berücksichtigt werden, können im nächsten Schritt mögliche Visualisierungen entwickelt und getestet werden, um die Anwendbarkeit neuartiger Darstellungen im Journalismus zu überprüfen.

Danksagung

An dieser Stelle möchte ich mich bei Dr. Tatiana von Landesberger sowie Prof. Dr. Chris Biemann bedanken, die mit ihrem konstruktiven Feedback bei der Erstellung dieses Beitrags mitgewirkt haben. Besonderer Dank geht an Kathrin Ballweg, welche den Prozess der Anforderungsanalyse tatkräftig unterstützt hat.

Literaturverzeichnis

[Au15] Ausserhofer, J.: "Die Methode liegt im Code": Routinen und digitale Methoden im Datenjournalismus. In (A. Maireder et al., Hrsg.): Digitale Methoden in der Kommunikationswissenschaft, Berlin, 87-111, 2015.

[BG09] Borchardt, A.; Göthlich, S.: Erkenntnisgewinnung durch Fallstudien. In (S. Albers et al., Hrsg.): Methodik der empirischen Forschung. Gabler Verlag, 33-48, 2009.

[Br14] Brehmer, M. et al.: Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. IEEE Transactions on

Visualization and Computer Graphics, 20(12), 2271-2280, 2014.

[Br09] Briggs, M.: Data-driven journalism and digitizing your life. http://www.journalism20.com/blog/2009/07/14/data-driven-journalism-and-digitizing-your-life/, Stand: 29.04.2016.

[Do11] Dou, W. et al.: ParallelTopics: A probabilistic approach to exploring document collections. IEEE Conference on Visual Analytics Science and Technology, 2011.

[Fa14] Fahrer, U. et al.: Network of the Day: Interactive Visualization of Time-Dependent Entity Relation Networks. Darmstadt, Germany, Vision Modeling and Visualization Workshop, 2014.

[Ga13] Gangemi, A.: A Comparison of Knowledge Extraction Tools for the Semantic Web. In (P. Cimiano et al., Hrsg.): The Semantic Web: Semantics and Big Data: 10th International Conference. Springer Berlin Heidelberg, Berlin, Heidelberg, 351-366, 2013.

[GL10] Gläser, J.; Laudel, G.: Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen. 4. Aufl., Wiesbaden, VS-Verl., 2010.

[Gö13] Görg, C. et al.: Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw. IEEE Transactions on Visualization and Computer Graphics, 19(10), 1646-1663, 2013.

[Gy14] Gynnild, A.: Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. Journalism, 15(6), 713-730, 2014.

[Ha00] Havre, S. et al.: ThemeRiver: visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 8(1), 9-20, 2002.

[Jä15] Jänicke, S. et al.: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. in Eurographics Conference on Visualization, R. Borgo et al., Hrsg.: The Eurographics Association, 2015.

[Ka06] Kang, H. et al.: NetLens: Iterative Exploration of Content-Actor Network Data. IEEE Symposium On Visual Analytics Science And Technology, 2006.

[KLB14] Kochtchi, A. et al.: Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles. Comput. Graph. Forum, 33(3), 211-220, 2014.

[Le05] Lee, B. et al.: Understanding research trends in conferences using paperLens. in Extended Abstracts on Human Factors in Computing Systems. Portland, OR, USA: ACM, 1969-1972, 2005.

[Lu07] Ludwig, J.: Investigativer Journalismus. 2. überarb. Aufl., Konstanz, UVK-Verl.-Ges., 2007.

[Ma10] Matzat, L.: Data Driven Journalism: Versuch einer Definition. http://datenjournalist.de/data-driven-journalism-versuch-einer-definition/, Stand: 29.04.2016.

[MN09] Meuser, M.; Nagel, U.: Das Experteninterview - konzeptionelle Grundlagen und methodische Anlage. In (S. Pickel et al., Hrsg.): Methoden der vergleichenden Politik- und Sozialwissenschaft. Verlag für Sozialwissenschaften, 465-479, 2009.

[RHR16] Roberts, J. C. et al.: Sketching Designs Using the Five Design-Sheet Methodology. IEEE Transactions on Visualization and Computer Graphics, 22(1), 419-428, 2016.

[SMM12] Sedlmair, M. et al.: Design Study Methodology: Reflections from the Trenches and the Stacks. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2431-2440, 2012.

[SG06] Stellman, A.; Greene, J.: Applied Software Project Management. O'Reilly Media, 2006.

[Su13] Sun, G.-D. et al.: A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges. Journal of Computer Science and Technology, 28(5), 852-867, 2013.

[Su16] Sun, M. et al.: BiSet: Semantic Edge Bundling with Biclusters for Sensemaking. IEEE Transactions on Visualization and Computer Graphics, 22(1), 310-319, 2016.

[UK15] Uskali, Turo; Kuutti, Heikki: Models and Streams of Data Journalism. The Journal of Media Innovations, 77-88, 2015.

[Zh13] Zhang, Z. et al.: The five Ws for information visualization with application to healthcare informatics. IEEE Transactions on Visualization and Computer Graphics, 19(11), 1895-1910, 2013.