Richtlinie zum „Software-Sprint“

Boromeo – Erkenntnisse aus Dokumentensammlungen sichtbar machen

Schlussbericht

Zuwendungsempfänger:
Simon Wörpel

*Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen **01IS19S19** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.*

Stand: 29. September 2020

Kurze Darstellung der Aufgabenstellung und Motivation

Die Diskussion um “Vertrauen in die Medien” oder “alternative Fakten”, um nur zwei Schlagworte zu nennen, ist gerade in den vergangenen Monaten der weltweiten Corona-Pandemie wieder erneut entflammt. Doch schon länger sind Medien, und damit ihre Rechercheur*innen, einem wachsenden Vertrauensanspruch seitens der Konsument*innen ausgesetzt. Erkenntnisse sollen und müssen transparent und nachvollziehbar sein. Nur so entsteht Glaubwürdigkeit. Eine Möglichkeit, um diesem geforderten Anspruch gerecht zu werden, kann darin bestehen, originale Quellen, auf die sich Fakten & Aussagen stützen, ebenfalls zu veröffentlichen. Sofern es der Quellenschutz zulässt.

Damit das überhaupt möglich ist, müssen Recherche-Teams, egal ob in Medien oder anderen Organisationen, auch intern einige Herausforderungen meistern: Wie generieren wir Erkenntnisse aus großen Dokument-Sammlungen? Wie überführen wir diese in strukturierte Fakten? Wie stellen wir sicher, dass wir unsere Erkenntnisse bzw. ihre Quellen später wiederfinden – zum Beispiel für eine neue Recherche? Wie können wir kollaborativ aber möglichst sicher zusammenarbeiten? Wie können wir unsere eigenen, auf Dokumenten basierenden Recherchen, effizient fakt-checken?

Das Vorhaben „Boromeo“ sollte diesen Überlegungen die nötigen technischen Lösungen bieten. Geplant war eine Plattform, die es Teams ermöglicht, an Dokumenten zu arbeiten und die daraus gewonnenen Erkenntnisse kollaborativ zu teilen sowie zu veröffentlichen. “Boromeo” sollte sich dabei auf einen möglichst einfachen Recherche-Workflow, die Verschneidung verschiedener Datenquellen und ein modulares und offenes System konzentrieren, um mit anderen, bestehenden Lösungen Daten & Erkenntnisse auszutauschen und zusammen zu arbeiten.

Im Verlauf des Vorhabens wurde deutlich, dass diese Ziele zu allgemein und zu umfangreich gesteckt waren, und es wurde sich auf einen konkreten Anwendungsfall konzentriert: Dem Annotieren von PDF-Dokumenten. Diese Prototyp-Lösung bekam den Namen „Annotate & Chill“ und wird im weiteren Verlauf dieses Berichts so benannt. Dennoch wurden weitere Aspekte des Vorhabens bearbeitet, worauf später noch eingegangen wird.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Die Zielgruppen des Vorhabens sind vielfältig. Primäre Zielgruppe sind Journalist*innen, die mit großen Dokumenten-Sammlungen, sogenannten Leaks, arbeiten. Das können interne Unterlagen, Ermittlungsakten aus einem Verfahren oder eine Sammlung von E-Mails sein. Diese Zielgruppe ergab sich aus dem bisherigen Betätigungsfeld des Projektteilnehmers, dem investigativen Journalismus. Jedoch sind alle Personengruppen, die in ihrer professionellen Arbeit mit solchen Dokumenten zu tun

haben, Teil der größeren Zielgruppe, da die technische Lösung für verschiedenste Arten von Dokumenten funktioniert.

Die nun in einem Prototypen veröffentlichte Lösung ermöglicht in erster Linie das Annotieren von PDF-Dokumenten sowie das (vom Dokument unabhängige, s.u.) Teilen dieser Daten mit anderen Mitarbeiter*innen. Dadurch können Recherche-Gruppen von der oben skizzierten Arbeitsweise profitieren, was ihrer Arbeit zu Transparenz und Vertrauen verhilft. Gleichzeitig ist das System auf die hohen Sicherheitsanforderungen solcher Teams ausgerichtet. Genau diese Ansprüche finden sich auch im Themenschwerpunkt „Engineering trust“ des Vorhabens wider.

Ausführliche Darstellung der Ergebnisse

Wie bereits berichtet wurde das Vorhaben auf einen konkreten Anwendungsfall, das Annotieren von Quell-Dokumenten, konkretisiert. Diese Überlegung wurde erst durch das Coaching, das im Rahmen der Begleitung durch die Open Knowledge Foundation ermöglicht wurde, deutlich und der Prozess hiervon begleitet. Es wurde zunächst analysiert, welche Lösungen es bereits gibt und daraus eine für das Vorhaben entwickelt, dass drei Voraussetzungen des oben skizzierten Nutzungszenarios erfüllt, die bisher von bestehenden Lösungen noch nicht angeboten wurden: Erstens, Annotierung kann lokal und offline auf dem Arbeitsgerät des Mitarbeitenden stattfinden ohne in einer Cloud gespeichert zu werden. Zweitens, Annotierungen sind getrennt von den Quell-Dokumenten und können unabhängig davon verbreitet und geteilt werden. Und drittens gibt es kein festgelegtes Format für Annotationen, es kann je nach Anwendungsfall und verwendeter Dritt-Software flexibel gestaltet werden.

Dies wird durch „Annotate & Chill“ dadurch gelöst, dass pro annotiertem Dokument Dateien im CSV-Format generiert werden, die unabhängig vom Quell-Dokument über sichere Kommunikationswege (beispielsweise verschlüsselte E-Mails) geteilt werden kann. Besitzen Mitarbeitende eines Recherche-Teams das selbe Dokument, müssen sie nur die Datei mit den Annotationen untereinander teilen und können diese auf ihrem Gerät mit dem Quell-Dokument verbinden. Das erfüllt die gestellten Sicherheitsanforderungen solcher Teams.

Die Anwendung funktioniert offline, also lokal auf dem Rechner. Es werden PDF-Dokumente importiert, die dann mit Annotationen versehen werden. Diese können über das Interface exportiert, auf einem anderen Gerät importiert und weiter bearbeitet werden. Die eindeutige Identifikation von Dokumenten funktioniert über deren Prüfsumme. Aus jeder Datei lässt sich diese Zeichenkette errstellen, die als eindeutige ID funktioniert. So können Annotationen einer Datei immer zugeordnet werden, egal ob das Dokument sich im einen Fall auf dem Laptop einer Person oder in einem anderen Fall in einer Online-Datenbank befindet.

Die Definition, auf welche Stelle sich die Annotation innerhalb eines Dokuments bezieht, wird als ein Rechteck auf der betreffenden Seite angegeben, das die Rechercheur*innen mittels des Interfaces mit der Maus zeichnen. So ist sichergestellt, dass sowohl PDF-Dokumente, die ein sog. „Text-Layer“ (also

extrahierten, maschinenlesbaren Text) enthalten, annotiert werden können, als auch Dokumente ohne Text (eingescannte PDFs) oder Bilder. Auf die Möglichkeit, Text-Passagen zu markieren und darauf basierend Annotationen zu erstellen, wurde aus Gründen der Einheitlich verzichtet, allerdings besteht im Server-Backend (s.u.) diese technische Möglichkeit.

Das Interface, sozusagen der „Client“ von „Annotate & Chill“, läuft derzeit im lokalen Browser des Nutzenden. Es ist jedoch geplant, diese Lösung als eigenständige Applikation für Windows, Mac und Linux anzubieten, damit sie einfacher zu nutzen ist. Auch wenn vordergründig damit der dezentrale, lokale (und somit sehr sichere) Anwendungsfall abgedeckt ist, wurde darüber hinaus auch eine Server-Anwendung entwickelt, die kollaboratives Arbeiten in der Cloud ermöglicht. Diese kann von Teams selbst betrieben werden, um an die jeweiligen Sicherheitsansprüche eines Projekts angepasst zu werden (z.B. Erreichbarkeit nur über einen VPN-Zugang). Mittels dieser Plattform werden Annotationen mit eindeutigen URLs versehen und können der Allgemeinheit zugänglich gemacht werden. Eine Datenbank hält vor, an welchen öffentlichen Orten im Netz die Quell-Dokumente verfügbar sind und verlinkt dorthin. Dieser Server-Teil von „Annotate & Chill“ soll demnächst als ergänzendes Angebot zum Prototypen veröffentlicht werden.

Somit konnte das hauptsächliche Anliegen, eine Lösung zu entwickeln um Erkenntnisse aus großen Datensammlungen strukturiert & kollaborativ zu bearbeiten, umgesetzt werden, wenn auch in etwas abgewandelter Form als zunächst in den Meilensteinen des Vorhabens skizziert. So wurde wie bereits erwähnt durch das angebotene Coaching das ursprüngliche Produkt, eine größere, allgemeine Rechercheplattform, in kleinere Teil-Lösungen unterteilt. Dies war zum Einen durch den Umfang dieses Vorhabens, der den Projekt-Rahmen deutlich gesprengt hätte, bedingt, und zum Anderen vor allem durch sinnhafte Überlegungen geprägt:

So gibt es bereits sehr gut funktionierende Komplett-Lösungen für Dokumenten-Recherchen, wie zum Beispiel “Datashare” vom International Consortium of Investigative Journalists (ICIJ) und insbesondere “Aleph” vom Organized Crime and Corruption Project (OCCRP). Beides sind Open-Source-Lösungen, die auch von anderen Newsrooms und Organisationen eingesetzt und mitentwickelt werden. Mit dem “follow the money”-Werkzeugkasten als Teil des Aleph-Systems ist es möglich, die unstrukturierte, reale Welt in maschineller Form so zu modellieren, sodass sie auch ein Computer verstehen kann.

Das hier beschriebene Vorhaben wurde dementsprechend konkretisiert, um den bestehenden Lösungen bisher nicht berücksichtigte Features (Annotation, getrennt vom Dokument) hinzuzufügen. „Annotate & Chill“ könnte in die bestehenden Plattformen integriert werden (s.u.) und ist glücklicherweise keine weitere „All-in-One“-Lösung eines in sich geschlossenen Systems geworden.

Als weiterer Meilenstein wurde eine Materialsammlung von Schulungsmaterial skizziert. Auch wenn durch die Grundsatz-Recherche während des Vorhabens hierzu bereits sehr viel Dokumentation gesammelt und aufgebaut werden konnte, reichte die Zeit nicht aus, dieses auch in gewinnbringender Form für andere Rechercheur*innen zu veröffentlichen. Dennoch wurde deutlich, dass die

Informations- bzw. Schulungslandschaft in diesem Bereich immer noch sehr wenig anbietet, vor allem für Einsteiger, sodass es ein großes (auch persönliches) Anliegen des Projektteilnehmers ist, diese Inhalte noch im Nachgang des Förderzeitraums auf einer Plattform zu veröffentlichen.

Insgesamt hat insbesondere die Coaching-Begleitung der Open Knowledge Foundation dazu beigetragen, das „übergroße“ Plattform-Vorhaben in kleinere, machbare Teil-Lösungen herunterzubrechen, von der die wichtigste, „Annotate & Chill“, bereits als Prototyp fertiggestellt und veröffentlicht wurde. Weitere Teilbausteine sind bereits in Arbeit oder sollen zeitnah angegangen und veröffentlicht werde, worauf weiter unten eingegangen wird.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Die oben skizzierten Zielgruppen können mit „Annotate & Chill“ nun auf einfache und unkomplizierte Weise Dokumente annotieren. Dies kann, wie beschrieben, kollaborativ und sicher geschehen und benötigt keine Server-Installation der Software. Die einfache Trennung von Quell-Dokument und Annotationen war ohne Cloud-Lösung auf diese Weise vorher nicht möglich.

Auch wenn „Annotate & Chill“ eine eigene Lösung ist, kann der technische Unterbau dank der Open-Source-Veröffentlichung in andere Plattformen integriert werden. Hierbei ist jede Plattform denkbar, die einen „PDF-Reader“ integriert hat, wie zum Beispiel „Aleph“ (http://alephdata.org/) oder Projekte der Open Knowledge Foundation wie „FragDenStaat“ (https://fragdenstaat.de/) oder „SehrGutachten“ (https://sehrgutachten.de/). Dazu müssten nur die technischen Komponenten, die die Dokumente auf der Website anzeigen, durch die zugrunde liegende Lösung von „Annotate & Chill“ ausgetauscht werden, die dann das Annotieren dieser Dokumente ermöglichen würde.

Außerdem soll, wie bereits erwähnt, die Client-Software als „Applikation“ ausgeliefert werden, die einfach auf den Arbeitsgeräten der Rechercheur*innen installiert werden kann. Hierzu soll in Kürze die Client-Anwendung in das „Electron“-Framework (https://www.electronjs.org/) integriert werden, was eine einfache technische Lösung für die Bereitstellungen von browserbasierten Webanwendungen als installierbare „System-Programme“ für Windows, Mac und Linux bietet.

Im Rahmen der Arbeit ist ein weiteres mögliches Projekt entstanden, das auch in Teilen dem ursprünglichen Vorhaben-Gedanken von „Boromeo“ entspricht: Eine Online-Plattform mit aufbereiteten und kuratieren Datensätzen, die einfach auffindbar sind und von Rechercheur*innen heruntergeladen und in ihre Recherche-Systeme integriert werden können. An dieser Lösung wurde als Teil des Projekt-Vorhabens exemplarisch bereits mit dem Daten-Team des „Organized Crime and Corruptiuon Projects“ (OCCRP) gearbeitet. Diese Plattform soll noch in diesem Jahr 2020 veröffentlicht werden.

Darüber hinaus wurde im Projektverlauf durch die intensive Arbeit an grundsätzlichen Themen zu Dokumenten-Verarbeitung und Verwaltung und das Studium der verschiedenen bestehenden

technischen Plattform-Lösungen ein Prototyp entwickelt, der das stillgelegte Projekt „Kleine Anfragen“ (https://kleineanfragen.de/) durch eine neue Lösung weiterführen könnte. Hierzu werden bereits Gespräche mit den entsprechenden Stellen geführt, um aus den gewonnenen Erkenntnissen dieser Projektlaufzeit die KleineAnfragen-Plattform neu zu entwickeln und weiterleben zu lassen.

Insgesamt hat nicht nur die technische Entwicklungsarbeit am Prototypen „Annotate & Chill“, an der neue Technologien ausprobiert werden konnten, den Projektteilnehmer fachlich weitergebildet, sondern vor allem die damit einhergehende Grundsatz-Recherche zu bestehenden (und immer noch fehlenden) Lösungen im weiten Feld der Dokumenten-Recherche. Der Projektteilnehmer konnte sich dadurch im Vorhabenverlauf intensiv persönlich und fachlich in den vielfältigen Themenbereichen fortbilden und hat nun ein deutlich größeren Wissens- und Erfahrungsschatz, der in der weiteren beruflichen Laufbahn sehr förderlich werden kann.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

In der ursprünglichen Vorhabens-Planung war angedacht, die Annotationen auf der bestehenden Lösung “hypothes.is” (https://web.hypothes.is/) aufbauend zu realisieren. Dieses Projekt stellt den Source-Code ebenfalls Open-Source zur Verfügung. Allerdings stellte sich bei genauerer Recherche heraus, dass diese technische Lösung nicht zu den identifizierten Nutzungs-Szenarien passt. So werden die Annotationen auf einem Server gespeichert und es wäre technisch sehr kompliziert gewesen, die Möglichkeit der offline-Bearbeitung in der Art, wie sie mit “Annotate & Chill” realisiert werden konnte, mit hilfe des “hypothes.is”-Systems zu realisieren. Aus einer Kosten-Nutzen-Abwägung wurde im weiteren Verlauf verzichtet, diesen Weg weiter zu verfolgen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Eine einfache Demo ist unter https://demo.annotate.cool zu erreichen. Hier können PDF-Dokumente annotiert werden und diese Annotationen im CSV-Format exportiert (und wieder importiert) werden. Auch wenn die Anwendung auf einem Server läuft, funktioniert sie technisch gesehen nur im Browser (Client) der Nutzer*innen, die PDF-Dokumente sowie die Annotationen werden niemals irgendwohin hochgeladen sondern nur lokal gespeichert.

Der Open-Source-Code dieser Anwendung ist unter https://github.com/simonwoerpel/annotate-and-chill/ erreichbar. Außerdem wurde im Rahmen der „Demo Week“ dieser Förder-Runde eine Website veröffentlicht, auf der das Vorhaben in einem Blog-Post konkret beschrieben und erklärt wurde: https://demoweek.prototypefund.de/projects/05-annotate-and-chill.html

Darüber hinaus soll in der nächsten Zeit unter dem ursprünglichen Projektnamen „Boromeo“ auf https://boromeo.org eine Präsentations-Seite der verschiedenen bearbeiteten Lösungen entstehen, die als Einstiegs-Portal dienen soll. Hier können sich User*innen über die technischen Angebote für das Recherchieren in Dokumenten informieren und die für sie passende Lösung, also unter anderem „Annotate & Chill“ samt Nutzerdokumentation, finden.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

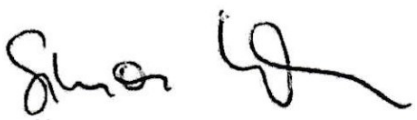
Zu Beginn des Projektes konnten nicht die geplanten Mittel abgerufen werden, da aufgrund der Corona-Pandemie und damit verbunder Mehraufwand in anderer freiberuflicher Arbeit des Projektteilnehmers weniger Arbeitszeit für die Fördermaßnahme aufgewendet werden konnte als zunächst geplant. Das wurde im weiteren Verlauf des Förderzeitraums wieder ausgeglichen. Deshalb wurde die zu Beginn veranschlagte Kostenplanung weitestgehend eingehalten.

Außerdem wurde der Arbeitsschwerpunkt auf die Entwicklung der technischen Lösung „Annotate & Chill“ gelegt und weniger auf die Plattform für Schulungsmaterial als ursprünglich geplant. Es wurde aber im Förderzeitraum viel dies betreffendes Material gesammelt und generiert, das noch im weiteren Verlauf der Arbeit an diesem Projekt veröffentlicht werden soll (s.o.)

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Eine verwandte Software-Lösung ist „Aleph“ vom Organized Crime and Corruption Project (OCCRP). Der Projektteilnehmer war bereits vor dem Förderzeitraum mit diesem Angebot vertraut. Ursprünglich arbeitete er an einer alternativen Lösung zu Aleph, die ebenfalls alle Anwendungsfälle vom Einlesen verschiedener Dateiformate, Analysieren der vorhandenen und generierten Daten und das kollaborative Recherchieren darin abbilden sollte. Es gab konkrete Feature-Ideen, die sich von Aleph und dem dort vorhandenen Angebot unterschieden hätten. Jedoch wurde die Software Aleph insbesondere in der ersten Jahreshälfte 2020 von dem dortigen betreuenden Team deutlich weiterentwickelt und um wichtige Features erweitert, sodass der Projektteilnehmer sich entschloss, die Arbeit an einer alternativen, „All-in-One“-Lösung einzustellen und sich stattdessen auf eine zugespitzte, modulare Lösung zu konzentrieren, aus der „Annotate & Chill“ entstanden ist.

Berlin, 30. September 2020



Simon Wörpel