Richtlinie zum „Software-Sprint“

Project “Follow the Grant”, gefördert als “Conflictbase”

Schlussbericht

Zuwendungsempfänger:
Boytchev Erdmann Scholze Wörpel GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen **01IS18S14** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Interessenkonflikte in Forschung und Medizin sind ein ernstes Problem für unsere Gesellschaft. Interessenskonflikte entstehen, wenn Forscher geldwerte Leistungen (etwa von Pharmafirmen) erhalten, was wiederum ihre Entscheidungen beeinflussen kann, oftmals unbewusst. Wissenschaftliche Studien belegen etwa, dass Zahlungen von Pharmafirmen mit einem Veränderten Verschreibungsverhalten einhergehen, das heißt, Ärzte können andere Medikamente verschreiben (für die sie gesponsert werden), als etwa für Patienten optimal sind. Oder Forscher äußern sich positiv zu Themen (z.B. Gentechnik) und werden gleichzeitig von Bayer (ehemals Monsanto) bezahlt.

Diese Interessenkonflikte sind für Bürger mehrheitlich nicht transparent und nachvollziehbar. Bisher gibt es nur wenige Ansätze für mehr Transparenz auf diesem Gebiet. Allerdings fordern wissenschaftliche Journals gewöhnlich von ihren Autoren, solche Interessenkonflikte bei der Publikation wissenschaftlicher Artikel offenzulegen. Doch diese Offenlegung möglicher Abhängigkeiten ist nicht in strukturierter Form vorhanden, ganz zu schweigen von journal- oder verlagsübergreifenden Informationen. So gibt es etwa für jeden Verlag, der diverse Journals herausgibt, andere Formen der Offenlegung, sowohl in Fließtext, als Fußnoten zu Veröffentlichungen, aber auch in strukturierter Form in den Meta-Daten eines wissenschaftlichen Papers.

Diese verschiedenen Informationsquellen der veröffentlichten Interessenskonflikte sollten in einer zentralen Datenbank gesammelt und für die Zielgruppe, Journalist*innen und interessierte Bürger*innen, verständlich und mit Kontext versehen auf einer Web-Plattform zugänglich gemacht werden. Dazu gehörte aus technischer Sicht in einem ersten Schritt die Sammlung dieser Informationen und in einem zweiten Schritt die technische Aufbereitung und Präsentation.

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

Das Projekt “Follow the Grant” (gefördert unter dem Arbeitstitel “Conflictbase”) trägt zu den Zielen von mindestens zwei der Schwerpunkte der Förderinitiative bei: “Diversität” sowie “Civic Tech”. Zu Ersterem (“Diversität”) ermöglichte das Projekt interdisziplinäre Zusammenarbeit eines diversen Teams, das – anders als klassische Software-Entwicklungsteams – neben Web-Entwicklern aus Journalisten und Daten-Experten bestand. Die unterschiedlichen Qualifikationen und erfahrungsbedingten Zugänge zum Thema der verschiedenen Team-Mitglieder, sowohl in technischer als auch inhaltlicher Hinsicht, ermöglichte einen Austausch über die gewohnten Horizonte hinweg und konnte wesentlich zur Qualität des Projekts beitragen.

Das Projekt “Follow the Grant” ermöglicht zudem für den Bereich “Civic Tech” der Zielgruppe aus Journalist*innen und Bürger*innen einen deutlich vereinfachten Zugang zu Daten, die zwar grundsätzlich der Öffentlichkeit zur Verfügung stehen (sofern die wissenschaftlichen Verlage sie in ihren Journals angeben), aber nicht zusammengeführt und durchsuchbar in einem frei zugänglichen Web-Projekt bisher verfügbar waren. Das Projekt ermöglicht somit nun der Zielgruppe durch den einfacheren Zugang auch die Möglichkeit der eigenen Meinungsbildung über die vorhandene oder nicht vorhandene Problematik der Interessenskonflikten in konkreten Fällen.

Ausführliche Darstellung der Ergebnisse

Die Software für “Follow the Grant” besteht aus mehreren einzelnen Komponenten, die unabhängig voneinander entwickelt wurden und genutzt werden können.

Zunächst wurde eine “Daten-Pipeline” entwickelt. Die Aufgabe dieser Software-Lösung ist, aus wissenschaftlichen Papers verschiedener Journals & Verlage Meta-Informationen sowie Angaben zu Interessenskonflikten zu extrahieren. Nach einer entsprechenden Aufbereitung können diese Daten in einer standardisierten Form in einen Such-Index (Elasticsearch) importiert werden. Dieser Teil der Software besteht aus einer Sammlung von Python-Skripten, die auf der Kommandozeile ausgeführt werden können. Das macht diesen Teil des Projekts sehr flexibel und skalierbar und erleichtert auch für andere, bisher nicht beteiligte Open-Source-Entwickler den Einstieg in die Mitarbeit.

Es wurden beim Extrahieren und Aufbereiten der Daten einige Muster erkannt, wie große Publisher ihre Webseiten gestalten. Das ermöglicht nun auch nicht-Programmierern, neue Komponenten zur “Daten-Pipeline” beizusteuern: Es braucht kein Python-Code geschrieben zu werden, sondern nur eine Definitions-Datei für einen neuen Verlag, die Informationen dazu beinhaltet, wo auf der Website dieses Verlags welche Informationen (Liste der Journals,

Papers, “klicke auf nächste Seite”) gefunden werden können. Allerdings können auch eigene Python-Module geschrieben werden, die konkret für einen bestimmten Verlag wirken, wenn dieser aus dem Standard-Muster fällt. Somit muss die Software nicht allein vom Followthegrant-Team gepflegt werden sondern kann von einer größeren interessierten Community aufgebaut und einfach erweitert werden.

Der zweite Teil der Software ist das “Backend”, das Daten für die Website bereitstellt. Das Backend ist eine simple Flask-App (python), die lediglich die Volltext-Suche durch den Elasticsearch-Index ermöglicht. Sie hat keine eigenen Seiten-Ansichten im Sinne einer klassischen Publikums-Website sondern gibt das Suchergebnis als JSON zurück, dass dann die Webseite nutzen kann. Diese Backend-App soll so simpel wie möglich bleiben.

Aus den Conflict of Interest Statements werden an dieser Stelle relevante Entitäten extrahiert. Da die Statement in ihrer Grammatik recht gleichförmig sind, können hierfür Methoden des Natural Language Processings eingesetzt werden. So können aus jedem Satz jeweils der Empfänger und der Geber von Zuwendungen festgestellt werden. Technisch wurde dies durch einen in Python implementierten Parse-Tree umgesetzt.

Der dritte Teil ist die eigentliche Website, die für die Zielgruppe zugänglich ist. Das Frontend dieser Website, eine Angular-Applikation, holt sich per Webservice-Anfrage die Daten zum vom Benutzer eingegebenen Suchbegriff. Zur Suche gelangt der Nutzer, indem er in der Navigationsleiste den Reiter Search auswählt. Hier befindet sich die Suchbox und unter dieser werden die Ergebnisse angezeigt. Der Suchbegriff kann der Name eines Autors, eines wissenschaftlichen Instituts, eines Themenfeldes oder Freitext sein. Die hierzu gefundenen Interessenskonflikte werden als Liste dargestellt. Ein Klick auf eines dieser Ergebnisse führt zur Detailansicht desselben. Hier sind die ausführlichen Information zu diesem Ergebnis sichtbar (Typ, Autor, Verlag, Paper, Wortlaut des angegebenen Interessenskonflikts, ..) Ebenfalls enthalten in der technischen Darstellung ist eine Übersichts-Seite mit der Beschreibung des Projekts.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Die Website followthegrant.org richtet sich an Journalisten, Experten, Wissenschaftler und interessierte Patienten und Bürger gleichermaßen. Sie ist trotz der Komplexität dahinter einfach zu bedienen. Neben allgemeinen Nutzern sind auch konkrete Kooperationspartner Teil der Zielgruppe: So sind mittelfristig Medienkooperationen mit Zeitungen oder Magazinen denkbar, die basierend auf dem Datenbestand bestimmte Themenfelder analysieren und in eigenen Veröffentlichungen thematisieren können. Erste Gespräche hierzu wurden bereits von der Projektleitung geführt.

Geplant sind in den nächsten Schritten eine umfassende Filterung der Ergebnisse nach relevanter Typisierung, nach Autoren, Magazinen, Journals und Themen. Die Weiterentwicklung des Frontend und der Nutzererfahrung sind dabei ständige Begleiter und werden iterativ immer

wieder neu in Frage gestellt und entsprechend verändert, erweitert oder aber auch wieder eingeschränkt.

Möglich wäre auch eine “intelligentere” Suche, indem die Conflict of Interest Statements durch Methoden aus dem Natural Language Processing weiter verarbeitet werden. So könnten Suchergebnisse automatisch übersichtlicher dargestellt werden, indem etwa vor der Trefferliste eine Übersicht aller Entitäten (also Autoren die Zuwendungen bekommen und Firmen, die solche geben) angezeigt wird.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Zunächst war geplant, innerhalb des Förderzeitraums auch Angaben zu möglichen Interessenskonflikten aus wissenschaftlichen Journals mit Bezahlschranke zu extrahieren. Dies erwies sich im Projektzeitraum aus technischer Sicht aufwendiger als zunächst veranschlagt. Das führte zu einer Prioritätenverschiebung zugunsten der Extrahierung aus Open-Access-Journals, sodass bis zum Projektende keine Journals mit Bezahlschranken in die Datenbank mit aufgenommen werden konnten. Zudem gab es auch rechtliche Unklarheiten im Zusammenhang mit dem “Scraping” von Inhalten aus nicht Open-Access-Journals, sodass darauf zunächst verzichtet wurde.

Bei der automatischen Extrahierung der Akteure ist die Auflösung von Entitäten (Entity Linking und Disambiguation) problematisch, da aus den Statements ohne Referenz teilweise keine Schlüsse gezogen werden können. So können etwa Autoren mit sehr häufigen Namen nicht eindeutig zugeordnet werden. Teilweise kann diese Problematik durch Einbezug der Institution, des Ortes und des Forschungsbereichs umgangen werden. Gerade Wissenschaftler wechseln allerdings oft die Institution und den Ort, an dem sie forschen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wir sind in Verhandlung mit dem Karlsruher Institut für Technologie, um eine Kooperation einzugehen und für dortige Wissenschaftler unsere Technologie zur Verfügung zu Stellen, damit zusammen eine Auswertung über Interessenkonflikte erfolgen kann. Palallel läuft auch ein Austuasch mit der Universitätsbibliothek Leipzig.

Für interessierte Software-Entwickler*innen ist der komplette Open-Source-Code für die verschiedenen Komponenten des Projekts (Scraper, API und Website) unter https://github.com/followthegrant unter einer freien Lizenz veröffentlicht.

Das Projekt wurde auf verschiedenen journalistischen Konferenzen präsentiert (Dataharvest 2018, Scicar 2018, European Science Journalism Conference 2018) und dort wurden Partner

für eine künftige Veröffentlichung gewonnen (z.B. ORF und Nature). Geplant für 2019 ist eine internationale Auswertung und veröffentlichung von Interessenkonflikten in verschiedenen Themenfeldern und Ländern. Dazu werden auch Bürgerinnen und Bürger nach Interessenkonflikten suchen können.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Zu Beginn des Projektes konnten nicht die geplanten Mittel abgerufen werden, da von den Projektteilnehmern weniger Arbeitszeit aufgewendet werden konnte als zunächst geplant. Das wurde im weiteren Verlauf des Förderzeitraums wieder ausgeglichen. Deshalb wurde die zu Beginn veranschlagte Kostenplanung weitestgehend eingehalten.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Neue Erkenntnisse erlangten die Teilnehmer zudem durch die beiden wahrgenommenen Workshops im Bereich User Experience Design, die im Rahmen des Projekts angeboten wurden. Diese fachliche Weiterbildung kann nun auch in weitere Projekten oder im beruflichen Arbeitsalltag angewendet werden.

Es ist ansonsten davon ausgzugehen, dass die entwickelte Technologie für anders geartete Auswertung von wissenschaftlichen Papers und Dataming von wissenschaftlicher Literatur eingesetzt werden kann.