FID-Netzwerk Semantische Technologien
(Organisation: FID Biodiversitätsforschung)

Zweiter Workshop zur Semantischen Extraktion von Informationen (SEI 2)

Di., 08.10.2024, als Videokonferenz (Zoom)
Beginn: 10.01 Uhr; Ende: 14.14 Uhr.

Begrüßung

Katrin Peikert begrüßt die Anwesenden (Teilnehmerliste im Anhang) und führt in die Thematik und in das Programm des heutigen 2. SEI-Workshops ein.

Kurzvorstellung der anwesenden FIDs *(insgesamt ca. 30 Min)*

In der Vorstellungsrunde werden die anwesenden FID-Projekte in alphabetischer Reihenfolge gebeten, sich kurz vorzustellen und aktuelle eigene Interessen und Services im Bereich semantischer Technologien darzustellen. Einige der vertretenen Projekte drücken aus, dass sie aus grundsätzlichem Interesse am Themenfeld teilnehmen, um auf einen aktuellen Stand zu kommen und Perspektiven für Kooperationen einzuschätzen. Eine Mehrheit der FID-Projekte hat selbst bereits Verfahren der semantischen Erschließung entwickelt und setzt diese im FID-Portal ein oder plant in absehbarer Zeit den Einsatz.

Vortrag “Beratung Text-and-Data-Mining” von Dr. Tillmann Dönicke (KfI)

inkl. Diskussion - 30 min

In seiner Präsentation zum TDM-Support des KfL beleuchtet Dr. Dönicke unter anderem Ergebnisse der Bedarfsermittlung; demnach haben bereits im Jahr 2022 über 80 % der FID-Projekte Bedarf an Unterstützungsdienstleistungen im Bereich TDM angemeldet. Eine weitere Umfrage Anfang 2024 erbrachte Erkenntnisse zu weiteren Details. Es wurden ein Sechs-Phasen-Modell (CRISP-DM-Modell) des Lebenszyklus eines TDM-Projekts vorgestellt und darauf aufbauend die Services des KfL erläutert und eingeordnet.

In der Diskussion zur Präsentation wurde geklärt, dass sich die Beratungsangebote direkt an die FID wenden; dagegen sei es nicht vorgesehen, dass die FID einzelne Forscher oder Forschergruppen an den TDM-Support weiterverweisen. Da in der TDM-Beratung auch urheber- und lizenzrechtliche Fragen relevant sein können, decke das Team im KfL diese Themen mit einer Juristin ab.

Vortrag “UCE: A Semantic Search Tool for Navigating UIMA-Annotated Corpora” von Kevin Bönisch (TextTechnologyLab / BIOfid) *inkl. Diskussion - 30 min*

Herr Bönisch reichert zunächst einige Darstellungen zu texttechnologischen Verfahren mit Angaben zum Bedarf an Ressourcen bzw. Speicherplatz an, um dann das Docker Unified UIMA Interface (DUUI) vorzustellen, welches in der Lage ist, die zuvor aufgestellten Probleme zu bewältigen und zu lösen. Die Inhalte annotierter Korpora können dann im Unified Corpus Explorer (UCE) tiefgehend analysiert, durchsucht und visualisiert werden. Der erwartete Leistungsumfang des UCE wurde anhand einer Live-Demo mit BIOfid-Daten vorgestellt; eine Nutzbarkeit für andere Fächer werde angestrebt.

In der Diskussion konnte geklärt werden, dass der UCE derzeit eine Evaluation durchlaufe und noch nicht verfügbar sei; später erfolge die Implementation in BIOfid sowie die Bereitstellung in einem GitHub-Repositorium zur Nachnutzung, woran mehrere im Workshop vertretene FIDs ihr Interesse ausdrückten. Weiterhin diskutiert werden Aspekte der Anpassbarkeit des UCE durch Integration von Fachvokabularen anderer Fächer.

Vortrag “Narrative Service” von Dr. Hermann Kroll (FID Pharmazie) *inkl. Diskussion - 30 min*

In seiner Präsentation zum Narrativen Service ordnet Dr. Kroll den Service in den Gesamtkontext der Angebote des FID Pharmazie ein und erläutert das Verhältnis zu den Drug Overviews. Der dem Narrativen Service zugrundeliegende Ansatz der graphenartigen Darstellung komme den fachlichen Arbeitsweisen entgegen. Eine Knowledge Extraction Toolbox, die vom FID Pharmazie entwickelt und gemeinsam mit dem FID Politikwissenschaften erprobt wurde, ermögliche fachunabhängige Nachnutzungen für die semantische Extraktion von Informationen.

In der Diskussion zur Präsentation wurde die Erkennung von Named Entities bzw. Konzepten thematisiert; diese erfolge nicht über Machine Learning, sondern über hochentwickelte Vokabulare, die in der Pharmazie seit langem etabliert seien. Eine Übertragung des Ansatzes auf die Politikwissenschaften zeigte, dass dort geeignete Vokabulare weniger gut verfügbar sind, so dass andere Verfahren, bspw. unter Verwendung von Word Embeddings in Frage kommen. Insgesamt erfordert der Ansatz des FID Pharmazie gute Vokabulare und klare Beschreibungen von Beziehungen zwischen Konzepten.

45 min Mittagspause

Vortrag “PhilFinder” von Nils Geißler (FID Philosophie) *inkl. Diskussion - 30 min*

Bei der Präsentation des PhilFinder zeigt Herr Geißler die Implementierung einer Knowledge Box für bekannte Entitäten im FID-Portal. Als Datengrundlagen werden sowohl Wikidata als auch die GND herangezogen; erläutert wurden auch Argumente für oder gegen die Verwendung bestimmter Identifikatoren. Daraufhin wurde PhiWiki

demonstriert, basierend auf dem durch NFDI4Culture bereitgestellten Wikibase4Research. Die im Rahmen eines Pilotprojekts generierten Daten könnten potenziell in Wikidata und GND einfließen.

In der Diskussion wurde unter anderem ein möglicher Datenfluss von Wikidata in die GND angesprochen. Dabei sei neben der Frage der Wünschbarkeit einer Anreicherung der als sehr zuverlässig angesehenen GND-Daten auch die Machbarkeit halbautomatischer Verfahren noch nicht geklärt; des Weiteren stehe auch eine Diskussion solcher Vorhaben mit dem GND-Ausschuss noch aus. Betreffend eine Frage nach einem Vergleich von Wikidata und GND konnte auf einen kommenden Konferenzbeitrag verwiesen werden.

Vortrag “Wissensgraph Bildung” von Marcel Jungmann (FID Erziehungswissenschaften & Bildungsforschung) *inkl. Diskussion - 30 min*

In seiner Präsentation zum Wissensgraphen Bildung ging Herr Jungmann von Literaturmetadaten als dem zentralen Gegenstand aus. Angestrebt werden Verknüpfungen von Publikationen mit Personen unter Nutzung von GND-ID und ORCID. In den Wissensgraphen fließen Datenbanken bzw. Repositorien aus dem "Such- und Nachweisraum" des FID ein und werden mit externen Informationsquellen verknüpft und angereichert. Bei der geplanten Implementierung von Personen-Seiten stelle sich teilweise die Herausforderung der nachträgliche Kuratierung von automatisch aggregierten Profildaten aus GND.

Die Diskussion zur Präsentation thematisierte unter anderem die Eignung verschiedener Datenquellen bzw. Personen-IDs. Die Zuhörer erfragten zudem, welche Ergebnisse oder Services bereits in das Portal des FID Erziehungswissenschaften & Bildungsforschung integriert seien: Bisher sind diese Ergebnisse noch nicht im Portal des FID Erziehungswissenschaften & Bildungsforschung implementiert, da sich das Vorhaben noch in der Entwicklung befindet.

Abschließende Diskussion *- 30 min*

Nach Abschluss des Vortragsprogramms konnten allgemeine Aspekte der Zusammenarbeit im FID-Netzwerk Semantische Technologien diskutiert werden. Hierzu waren mit den Bordmitteln der Videoplattform Zoom Umfragen unter den Teilnehmenden vorbereitet worden. Eine breite Mehrheit von etwa 85 % der Anwesenden bekundete ein Interesse, themenspezifisch zusammenzuarbeiten oder sich in themenspezifischen AGs auszutauschen. Eine Themensammlung auf einem virtuellen Whiteboard ergab folgende sechs Punkte (Reihenfolge nicht wertend, gemäß zufälliger Anordnung auf dem Whiteboard): "Interaktive Visualisierung von (RDF-)Graphen"; "ORCID"; "Erschließung von Volltexten"; "Retrieval Augmented Generation (RAG) mit SPARQL Endpoint/RDF, am liebsten in Python"; "GND Datenimport"; "Best Practices Kuratierung". Im weiteren Verlauf der Diskussion kam es zu einer Abstimmung darüber, ob eine Bearbeitung der genannten Themen eher in Kleingruppen oder in einem Gesamttreffen des Netzwerks sinnvoll sei. Eine breite Mehrheit von etwa 70 % sprach sich für eine Bearbeitung solcher Themen im Rahmen von Gesamttreffen aus. Zu der Frage, ob das nächste Treffen des Netz-

werks erst in einem Jahr oder bereits in einem halben Jahr stattfinden solle, ergab sich ein einigermaßen ausgeglichenes Bild mit einer leichten Mehrheit für "in einem Jahr". Auf diesem Ergebnis basierend bildete sich der Konsens, das nächste Treffen des Netzwerks in einem dreiviertel Jahr folgen zu lassen.

Mehrere diskutierende Personen plädierten für eine stärkere Einbeziehung von Ansätzen und Akteuren aus dem NFDI-Kontext. Der das Netzwerk organisierende FID (BIOfid) bekundete die Absicht, dies in der Gestaltung des Programms für das kommende Netzwerk-Treffen zu berücksichtigen.

In einem Fazit der heutigen Veranstaltung stellte Herr Kasperek fest, dass durch die Vorträge und Diskussionen die Breite des Themas und die vielfältige Relevanz von semantischen Technologien für FIDs sehr deutlich geworden sei. Semantische Technologien werden in FID-Kontexten sowohl auf Volltexte oder Abstracts als auch auf bibliographische Daten im engeren Sinne sowie auf Normdaten angewendet. Zudem sei klar geworden, dass je nach Fach verschiedene Herangehensweisen eine Berechtigung haben - je nachdem, wo die jeweiligen fachlichen Notwendigkeiten und Bedarfe liegen.

Im Hinblick auf die Terminplanung für das kommende Netzwerk-Treffen wurde der Wunsch formuliert, ein Zeitfenster zu wählen, das entweder den Vormittag oder den Nachmittag belegt, und sich dabei nicht über die Mittagszeit erstreckt. Da das nächste Treffen wiederum als virtuelles Treffen vorgesehen ist und da keine Zeitaufwände für die An- oder Abreise am Morgen bzw. Nachmittag entstehen, soll dieser Vorschlag umgesetzt werden.

Teilnehmerliste

Altmeier, Nicole	KfL
Blume, Patricia	FID Media
Bönisch, Kevin	FID Biodiversitätsforschung (BIOfid)
Brandt, Olaf	UB Tübingen / Technik-Board
Cramme, Stefan	FID Erziehungswissenschaft und Bildungsforschung
Dias, Pedro Henrique	FID Biodiversitätsforschung (BIOfid)
Dönicke, Tillmann	KfL
Faßnacht, Martin	FID Theologie
Flietner, Florian	FID Nord
Franke-Maier, Michael	FID AAC
Fuchs, Matthias	FID Move
Geißler, Nils	FID Philosophie
Grüter, Doris	FID Romanistik
Hamann, Olaf	FID Slawistik
Hampf, Yannik	FID Philosophie

Ho, Brent	FID Asien (CrossAsia)
Israel, Holger	FID Physik
Jungmann, Marcel	FID Erziehungswissenschaft und Bildungsforschung
Kasperek, Gerwin	FID Biodiversitätsforschung (BIOfid)
Kasprzik, Anna	FID Wirtschaftswissenschaften (EconBiz)
Kim, Timotheus	FID Theologie
Koch, Franziska	FID Politikwissenschaft
Kroll, Hermann	FID Pharmazie
Meister, Janina	FID Kriminologie
Nötzel, Swantje	FID Nahost
Ohms, Jannis	FID Pharmazie
Peikert, Katrin	FID Biodiversitätsforschung (BIOfid)
Porath, Antina	KfL
Renner-Westermann, Heike	FID Linguistik
Riek, Ilona	FID Benelux
Sarayeva, Asya	FID Slawistik
Schepers, Leon	FID Romanistik
Schmidt, Marie-Luise	FID Jüdische Studien
Schnabel, Benjamin	FID Jüdische Studien
Ulrich, Ivo	FID Slawistik
Vogel, Ivo	FID intRecht
Vrdoljak, Ivana	FID Linguistik
Weber, Tobias	FID Linguistik
Wulle, Stefan	FID Pharmazie