Narrativer Informationszugriff interdisziplinär

Chancen und Herausforderungen für Fachinformationsdienste

Wolf-Tilo Balke, Technische Universität Braunschweig, Institut für Informationssysteme
Hermann Kroll, Technische Universität Braunschweig, Institut für Informationssysteme
Florian Plötzky, Technische Universität Braunschweig, Institut für Informationssysteme
Julian Schenke, Technische Universität Braunschweig, Universitätsbibliothek

Zusammenfassung:
Digitale Bibliotheken implementieren typischerweise Schlüsselwort-basierte Zugriffspfade für ihre Kollektionen. Die Suche nach komplexen Informationszusammenhängen kann jedoch aufwändig werden, wenn Nutzende auf Schlüsselwörter beschränkt werden. Schließlich ist ein essenzieller Bestandteil des wissenschaftlichen Diskurses die Veröffentlichung von Wissen in stringenten Argumentationszusammenhängen. Besonders explorative Suchen erfordern Anfragen mit offenen Schlüsselwörtern, was angesichts der stark ansteigenden Zahl von wissenschaftlichen Veröffentlichungen zu einem erheblichen Aufwand während des Literaturscreenings führt. Eine Alternative zur herkömmlichen Suche mit Schlüsselwörtern könnte es sein, dass Nutzende ihren Informationsbedarf als Narrativ, d. h. als eine strukturierte Anfrage mit Bezug auf relevante Akteure und deren Interaktionen in ihrem eigentlichen Kontext zu formulieren. In enger Kooperation zwischen der Universitätsbibliothek und dem Institut für Informationssysteme der Technischen Universität Braunschweig wurde im Fachinformationsdienst (FID) Pharmazie (PubPharm) ein narrativer Informationszugriff für die Pharmazie (www.narrative.pubpharm.de) entwickelt und implementiert. In einer gemeinsamen Kooperation mit dem FID Politikwissenschaft (Pollux) wurde ein narrativer Informationszugriff für die Politikwissenschaft erprobt. Dabei wurden zunächst die entwickelten Extraktionsverfahren auf ihre Eignung für die Politikwissenschaft hin untersucht, ein narrativer Informationszugriff mittels semantischer Suche prototypisch entwickelt sowie am Beispiel europäischer Reden Chancen und Herausforderungen des Zugriffs erarbeitet. Die dabei gewonnenen, zentralen Erkenntnisse werden in diesem Artikel vorgestellt: Ein narrativer Informationszugriff für die Politikwissenschaft ist nützlich und hilfreich, jedoch aufgrund fehlender Vokabulare und geeigneter Extraktionsmethoden aufwändiger umzusetzen als für die Pharmazie.

Summary:
Digital libraries typically implement keyword-based access paths for their collections. However, searching for complex information needs can become challenging if users are limited to keyword searches. Eventually, an essential part of the scientific discourse is the publication of knowledge in stringent argumentation contexts. Especially explorative searches require the usage of open keywords, which leads to a considerable effort during literature screening in view of the rapidly increasing number of scientific publications. An alternative could be that we enable users to formulate their information need as a narrative, i.e., as a structured query about the relevant actors and their interactions in their actual context. In close cooperation between the University Library and the Institute for Information Systems of the Technical University of Braunschweig, narrative information access for pharmacy (www.narrative.pubpharm.de) has been developed and implemented in the Specialized Information Service (FID) Pharmacy (PubPharm). In joint cooperation with the FID Political Sciences (Pollux),

narrative information access for political sciences was evaluated. First, we examined the developed extraction methods for their suitability in political sciences. We prototypically implemented narrative information access through a semantic search and elaborated on the chances and challenges of the access in the example of European speeches. In this article, we present the central insights gained in that process: Narrative information access for political sciences is useful and helpful. However, due to a lack of concept vocabularies and suitable extraction methods, it is more complex to implement than for pharmacy.

Autorenidentifikation:
Balke, Wolf-Tilo: GND: http://d-nb.info/gnd/173565026,
ORCID: https://orcid.org/0000-0002-5443-1215
Kroll, Hermann: ORCID: https://orcid.org/0000-0001-9887-9276
Plötzky, Florian: ORCID: https://orcid.org/0000-0002-4112-3192
Schenke, Julian: GND: http://d-nb.info/gnd/1167969928,
ORCID: https://orcid.org/0000-0002-3584-2027

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5962

Schlagwörter: Information Retrieval; Narrativer Informationszugriff; Discovery System; Fachinformationsdienst

1. Einleitung

Digitale Bibliotheken implementieren typischerweise Schlüsselwort-basierte Zugriffspfade für ihre Kollektionen. Nutzende müssen demnach ihren Informationsbedarf, egal wie komplex er auch ist, als Schlüsselwörter formulieren. Bei der Suche nach Fachliteratur in Disziplinen wie z.B. der Pharmazie spielen aber Mechanismen und Interaktionsmuster eine relevante Rolle, bspw. Beziehungen zwischen Wirkstoffen. Eine präzise Suche nach diesen Mustern kann sich als schwierig erweisen, weil die Beziehungen nur schlecht in Form von Schlüsselwörtern ausgedrückt werden können. Ebenso wäre eine Exploration der Literatur mittels Variablen-behafteter Muster wünschenswert: Welche Wirkstoffe wurden schon für eine Diabetes-Behandlung von Erwachsenen getestet? In welchen Arzneiformen kann ein bestimmter Wirkstoff für eine Behandlung einer Krankheit und Zielgruppe dargereicht werden? Solche Suchen lassen sich nur schwer mittels Schlüsselwörtern ausdrücken, da die Suchanfrage offen formuliert sein muss (z.B. ohne konkrete Krankheit/Zielgruppe), die Ergebnismengen damit groß werden und somit ein manuelles, aufwändiges Literaturscreening erfordern.

Strukturierte, und meist manuell kurierte, Datenbanken und Wissensbasen (oder auch Wissensgraphen) wie z.B. ChEMBL[1], DrugBank[2] oder Wikidata[3] kuratieren Wissen in einer strukturierten Form. Diese strukturierte Repräsentation ermöglicht es Nutzenden dann, ihre Anfrage auch als komplexes Muster von Interaktionen mittels Anfragesprachen wie SQL für relationale oder SPARQL für graphbasierte Datenbanken auszudrücken. Jedoch bringt dieser Ansatz zwei zentrale Probleme mit sich. Erstens: Das Kuratieren solcher Datenquellen ist aufwändig und kostenintensiv. Zweitens besteht der essenzielle Bestandteil des wissenschaftlichen Diskurses in der Veröffentlichung von Wissen in stringenten Argumentationszusammenhängen. Damit muss das kuratierte Wissen typischerweise aus Forschungsliteratur extrahiert werden. Dieses Wissen jedoch, das ursprünglich in kohärenten Argumentationslinien veröffentlicht wurde, muss bei der Extraktion auf eine vereinfachte, strukturierte Repräsentation z.B. als Subjekt-Prädikat-Objekt-Aussage (Wirkstoff behandelt Krankheit) reduziert werden. Hierbei kann der eigentliche Zusammenhang zwischen diesen Aussagen jedoch verloren gehen.

Eine Alternative zur herkömmlichen Suche mit Schlüsselwörtern könnte es sein, dass Nutzende ihren Informationsbedarf als Narrativ, d. h. als eine strukturierte Anfrage mit Bezug auf relevante Akteure und deren Interaktionen in ihrem eigentlichen Kontext formulieren. In der Pharmazie könnte ein Interaktionsmuster zwischen pharmazeutischen Konzepten ein solches Narrativ darstellen, z.B. die Behandlung einer Krankheit mit einer Kombination von verschiedenen Wirkstoffen. Ein solcher narrativer Informationszugriff ermöglicht auf einem entsprechend vorbereiteten Datenmaterial sowohl präzise Suchen als auch strukturierte Übersichten der eigentlichen Forschungsliteratur, indem Variablen-behaftete narrative Anfragen zugelassen werden (z.B. welche Wirkstoffe für eine Behandlung eingesetzt oder in welcher Arzneiform sie dargereicht werden).

In enger Kooperation zwischen der Universitätsbibliothek und dem Institut für Informationssysteme der Technischen Universität Braunschweig wurde vor einiger Zeit im Fachinformationsdienst (FID) Pharmazie (PubPharm) ein narrativer Informationszugriff für die Pharmazie entwickelt und implementiert.[4] In einer weiteren engen Kooperation zwischen PubPharm und dem FID Politikwissenschaft (Pollux) wurden nun die für die Pharmazie entwickelten Methoden auf ihre Übertragbarkeit in die Politikwissenschaft hin untersucht und evaluiert. Zudem wurde ein Prototyp entwickelt, der narrative Suchen im Kontext des zweiten Irakkriegs mit Bezug auf Wikipedia und Reden des europäischen Parlaments erlaubte. In diesem Artikel werden die zentralen Ergebnisse vorgestellt und zusammengefasst, beginnend mit dem narrativen Informationszugriff in der Pharmazie, gefolgt von der Generalisierbarkeit der eingesetzten Extraktionsverfahren auf die Politikwissenschaft sowie der Implementierung, Evaluierung und Diskussion des narrativen Zugriffs für die Politikwissenschaft.

1 ChEMBL, <https://www.ebi.ac.uk/chembl>, Stand: 10.07.2023.
2 DrugBank, <https://go.drugbank.com/>, Stand: 10.07.2023.
3 Vrandečić, Denny; Krötzsch, Markus: Wikidata. A Free Collaborative Knowledgebase, in: Communications of the ACM 57 (10), 2014, S. 78–85. Online: <https://dx.doi.org/10.1145/2629489>.
4 PubPharm, Narrative Service: <https://narrative.pubpharm.de/>, Stand: 10.11.2023.

2. Narrativer Informationszugriff in der Pharmazie

Der von PubPharm entwickelte und betriebene narrative Informationszugriff wurde andernorts bereits vorgestellt[5] sowie im Detail beschrieben[6]. Der Service basiert auf der zentralen Idee, dass Nutzende ihre Suchanfragen als Interaktionsmuster zwischen biomedizinischen Konzepten formulieren. Beispielsweise kann so nach Behandlungen in bestimmten Patient*innengruppen oder nach Wechselwirkungen zwischen verschiedenen Wirkstoffen gesucht werden. Ebenfalls werden Variablen in Suchen unterstützt, um explorative Anfragen zu ermöglichen, z.B. welche Wirkstoffe eine Krankheit behandeln oder wie diese dargereicht werden.

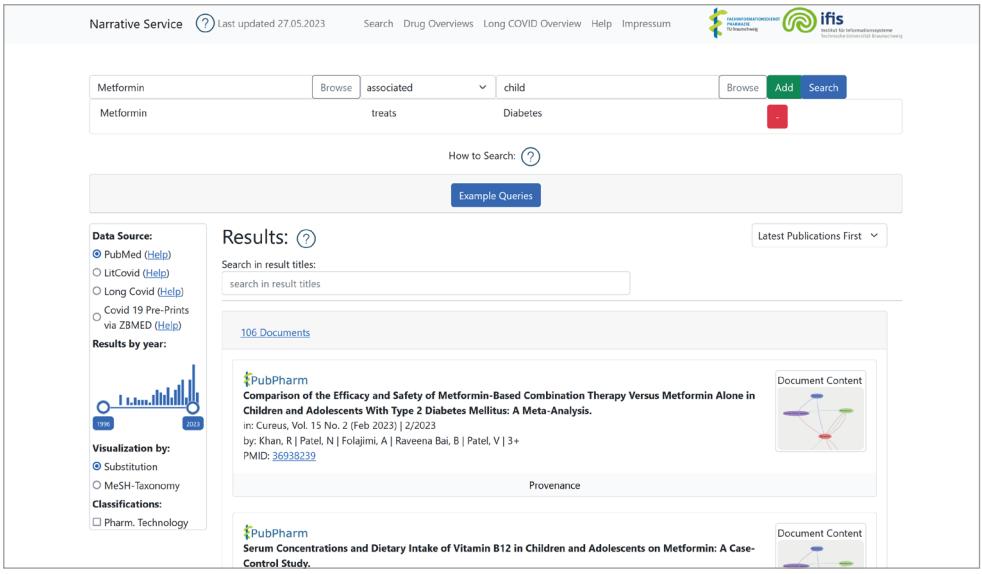

Abbildung 1: Benutzeroberfläche des Narrativen Service. Als Beispiel wurde eine Suchanfrage nach Diabetes-Behandlungen bei Kindern mit Metformin eingegeben.

Abbildung 1 zeigt die Benutzeroberfläche mit einer Beispielsuche. Treffer sind dann Veröffentlichungen, die das gesuchte Interaktionsmuster vollständig enthalten. Damit diese Anfragen entsprechend beantwortet werden können, wurde die Literatur einer Vorverarbeitung unterzogen. Zunächst wurden biomedizinische Konzepte (Wirkstoffe, Krankheiten, etc.) aus etablierten Taxonomien und Ontologien in den Texten ausgezeichnet sowie ihre Interaktionen extrahiert. Anschließend wurden die extrahierten Informationen in einem strukturierten Repositorium gespeichert. Jedes einzelne Dokument

5 Kroll, Hermann; Draheim, Christina: Narrative Information Access for a Precise and Structured Literature Search, in: o-bib 8 (4), 2021, S. 1–13. <https://doi.org/10.5282/o-bib/5730>.

6 Kroll, Hermann; Pirklbauer, Jan; Kalo, Jan-Christoph; Ruthmann, Johannes; Balke, Wolf-Tilo: A Discovery System for Narrative Query Graphs. Entity-Interaction-Aware Document Retrieval, in: International Journal on Digital Libraries, 2023. Online: <https://doi.org/10.1007/s00799-023-00356-3>.

wird dort als Interaktionsmuster zwischen Konzepten dargestellt. Abbildung 2 zeigt beispielhaft die Visualisierung eines Dokuments, wie man sie ausgehend von der Trefferliste abrufen kann.

Im narrativen Informationszugriff werden semantische Anfragemuster innerhalb von strikten Dokumentkontexten beantwortet. In anderen Worten: Um als Treffer ausgegeben zu werden, muss das gesuchte Muster vollständig innerhalb einer Veröffentlichung und damit innerhalb ihres Kontexts vorkommen (in dem vorgestellten Suchsystem in Abstracts). Mit Hilfe dieser Einschränkung wird sichergestellt, dass Muster mit zusammengehörigen und kontext-kompatiblen Informationen beantwortet werden, z.B. dass Patient*innen mit einer gesuchten Nebenwirkung auch wirklich mit dem angefragten Wirkstoff behandelt wurden. Genau eine solche Verbindung fehlt in traditionellen Datenbanken wie z.B. ChEMBL oder DrugBank, die bspw. nur Nebenwirkungen listen und die entsprechenden Kontextbedingungen nicht enthalten. Der narrative Informationszugriff ermöglicht demnach präzise und strukturierte Suchen mittels ausdrucksstarker Interaktionsmuster, die in Kontexten (hier Abstracts) beantwortet werden.

Der entwickelte Narrative Service und die pharmazeutische Annotationspipeline stehen mittlerweile als Open Source Projekte zur freien Verfügung[7].

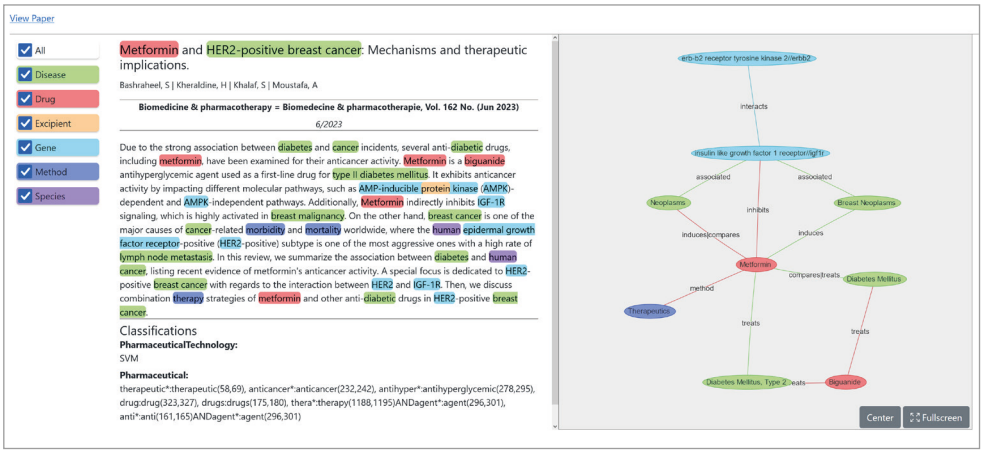

Abbildung 2: Repräsentation eines Dokuments: Auf der linken Seite sind Titel, Autoren, Journal sowie der Abstract einer Veröffentlichung dargestellt. Im Text erkannte pharmazeutische Konzepte sind farblich hervorgehoben. Auf der rechten Seite werden extrahierte Interaktionen als Graph visualisiert. Nutzende können bestimmte Konzepttypen optional abwählen.

7 Narrative Service Code: <https://github.com/HermannKroll/NarrativeIntelligence>, Stand: 30.11.2023; Pharm. Annotationspipeline: <https://github.com/HermannKroll/NarrativeAnnotation>, Stand: 30.11.2023.

2.1 Relevanz für die Politikwissenschaft

Das Vorhaben der Übertragung des narrativen Informationszugriffs auf die Politikwissenschaft dockt an einen Grundgedanken der Digital Humanities an, geistes- und sozialwissenschaftliche Fragestellungen unter Zuhilfenahme informationstechnologischer Methoden zu bearbeiten. Um die in den vergangenen Jahren zunehmend anfallenden großen Datenmengen zu bewältigen, werden hier zunehmend computergestützte Verfahren und digitale Ressourcen konsultiert.[8] Mit Blick auf die methodologischen Diskussionen in den Sozial- und Politikwissenschaften bedeutet das, Suchanfragen als fiktive bzw. hypothetische Weltaussagen in eine gegebene Dokumentenkollektion einspeisen zu können. Ergibt die Anfrage plausible Treffer, entspräche das einer „Proto-Diskursanalyse", d.h. einer vorbereitenden Suche nach Sinnzusammenhängen, oder ergäbe zumindest vorbereitende Erkenntnisse über die in der Dokumentenkollektion vorhandene semantische Struktur von Argumenten und Kausalbehauptungen. „Klassische" politikwissenschaftliche Diskurs- und Inhaltsanalysen werden i.d.R. intellektuell auf der Basis hermeneutischer Methoden vorgenommen; dieser Prozess ist in der praktischen Forschung oft so aufwändig und zeitintensiv, dass er einen großen Teil oder die Gesamtheit politikwissenschaftlicher Studiendesigns und Qualifikationsarbeiten dominiert. Ein entsprechendes narratives Suchsystem würde den Forschungsprozess signifikant beschleunigen. Allerdings schließt sich hier die Frage an, ob sich die in der Pharmazie eingesetzten Extraktionsverfahren eignen, um einen ähnlichen narrativen Informationszugriff auch in der Politikwissenschaft realisieren zu können. Die angestrebte Extraktion natürlicher Sprachinformationen, d.h. deren Re-Phrasierung als strukturierte narrative Information, soll die automatische Maschinenlesbarkeit semantischer Informationen ermöglichen und damit a) der Weiterentwicklung präziser und strukturierter Textrecherchedienste im Kontext (digitaler) Bibliotheken sowie möglicherweise b) der künftigen Erweiterung forschungsnaher Services dienen. Das Potenzial für die politikwissenschaftliche Forschung liegt hier in dem Vorhaben, die o. g. etablierten Auswertungsmethoden von Dokumentenkollektionen zu unterstützen.

3. Generalisierbarkeit der Extraktionsverfahren in der Politikwissenschaft

Ein gängiges Vorgehen bei der Extraktion von Wissen aus Texten ist der Einsatz von überwachten (supervised) Extraktionsmodellen. Dafür werden Trainingsdaten benötigt, also Beispiele, welche Interaktionen zwischen Konzepten in Sätzen vorliegen. Mit diesen Daten können dann Modelle trainiert bzw. angepasst (fine-tuned) werden. Gängige Beispiele dafür sind moderne Sprachmodelle (Language Models). Genügend und vor allem hoch qualitative Daten für jede Applikation zu sammeln, kann aufwändig und mit hohen Kosten verbunden sein.

8 Thaller, Manfred: Controversies Around the Digital Humanities. An Agenda, in: Historical Social Research 37 (3), 2012, S. 7–23, hier S. 7f. Online: <https://doi.org/10.12759/hsr.37.2012.3.7-23>.

In unserer vorherigen Arbeit haben wir deshalb sogenannte nahezu-unüberwachte (nearly-unsupervised) Extraktionsworkflows vorgeschlagen[9], Open Source veröffentlicht[10], im Detail untersucht[11] und für den Betrieb unseres Narrativen Service eingesetzt. Die Idee hierbei ist, Extraktionsverfahren wie z.B. Open Information Extraction einzusetzen, die keine domänen-spezifischen Trainingsdaten benötigten. Diese Methoden extrahieren strukturierte Aussagen basierend auf der grammatikalischen Struktur eines Satzes. Früher wurden diese Methoden regelbasiert implementiert, jedoch werden heutzutage auch Sprachmodelle (Language Models) für diese Aufgabe angelernt. Der Vorteil dieser Methoden ist es, dass sie direkt in einer Vielzahl von Domänen anwendbar sind und weder domänen-spezifische Trainingsdaten noch weitere Anpassungen benötigen. Eine extrahierte Aussage besteht typischerweise aus zwei Nominalphrasen und einer Verbphrase, z.B. <The drug metformin; treats; diabetes in patients>. Jedoch weisen diese Methoden den Nachteil auf, dass die extrahierten Aussagen normalerweise nicht kanonisiert (normalisiert) vorliegen: Nominalphrasen können unter anderem mehrere Konzepte aufweisen (z.B. diabetes in patients), dasselbe Konzept auf verschiedene, synonyme Weisen ausdrücken (z.B. the drug metformin oder nur metformin) oder dieselbe Interaktion mittels einer Vielzahl von Verbphrasen umschreiben. Die Idee der nahezu unüberwachten Workflows ist es nun, diese nicht kanonisierten Aussagen mit Hilfe von Konzept- und Relationsvokabularen zu filtern bzw. zu kanonisieren. Im obigen Beispiel könnte die erste Nominalphrase the drug metformin auf den Wirkstoff metformin reduziert werden und die zweite Phrase auf die Krankheit diabetes. Mit Hilfe eines weiteren, sogenannten Relationsvokabulars können auch verschiedene Verbphrasen wie treats, aids oder prevents zu präzisen Relationen wie z.B. treats zusammengeführt werden. In anderen Worten verlangen dieses Vorgehen damit keine spezifischen Trainingsdaten für eine Domäne, jedoch werden Vokabulare für Konzepte und Relationen benötigt, um die Extraktionen zu filtern bzw. auf relevante Extraktionen zu beschränken, um damit eine präzise Semantik sicherzustellen. Unser Argument ist, dass solche Vokabulare einfacher und kostengünstiger zu erstellen sind, als es die Akquise von Trainingsdaten sowie das Training geeigneter Modelle für jede einzelne Domäne erfordert.

Für PubPharms narrativen Informationszugriff konnten diese Extraktionsverfahren erfolgreich eingesetzt werden, genauer gesagt ein Pfad-basiertes Verfahren namens PathIE. Geeignete Vokabulare in Form von biomedizinischen Taxonomien und Ontologien sind in der pharmazeutischen Domäne weit verbreitet. Ein Beispiel dafür sind die von der U.S. Library of Medicine gepflegten Medical Subject Headings (MeSH)[12]. Diese Headings umfassen relevante Konzepte sowie zugehörige Beschreibungen, Kategorisierungen und zugehörige Listen von Synonymen. Zudem ist in der Pharmazie eine begrenzte Menge von Relationen von Interesse, z.B. kann ein Wirkstoff eine Krankheit behandeln oder diese als Nebenwirkung verursachen. PubPharm hat so in den letzten Jahren eine Text-Mining-Pipeline entwickelt, die biomedizinische Texte automatisiert in eine strukturierte Repräsentation überführt. Dabei werden relevante Konzepte in Texten detektiert und ihre Interaktionen extrahiert.

9 Kroll, Hermann; Pirklbauer, Jan; Balke, Wolf-Tilo: A Toolbox for the Nearly-Unsupervised Construction of Digital Library Knowledge Graphs, in: ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2021, S. 21–30. Online: <https://doi.org/10.1109/JCDL52503.2021.00014>.

10 Toolbox: <https://github.com/HermannKroll/KGExtractionToolbox>, Stand: 05.07.2023.

11 Kroll, Hermann; Pirklbauer, Jan; Plötzky, Florian; Balke, Wolf-Tilo: A Detailed Library Perspective on Nearly Unsupervised Information Extraction Workflows in Digital Libraries, in: International Journal on Digital Libraries, 2023. Online: <https://doi.org/10.1007/s00799-023-00368-z>.

12 Medical Subject Headings: <https://meshb.nlm.nih.gov/>, Stand: 05.07.2023.

Sowohl die Existenz von Konzeptvokabularen als auch die beschränkte Anzahl von relevanten Relationen sind besondere Merkmale der Pharmazie, was die Extraktion von strukturierten Aussagen erheblich vereinfacht. In einer gemeinsamen Arbeit mit dem FID Politikwissenschaft wurden die für PubPharm eingesetzten Extraktionsverfahren auf ihre Generalisierbarkeit überprüft. In der Politikwissenschaft existieren im Vergleich zu der Pharmazie keine derartigen und vor allem weitverbreiteten Konzeptvokabulare. Die Wortwahl, verwendete Begrifflichkeiten oder die allgemeine Beschreibung sind bspw. für eine Deutung bzw. Auslegung eines Texts entscheidend. Zusätzlich sind weitaus mehr verschiedene Relationen relevant, man bedenke bspw. allein die Menge an möglichen Verhältnissen zwischen Personen oder Organisationen. Die Adaption obiger Extraktionsverfahren auf die Politikwissenschaft ist damit schwierig.

In unserer Arbeit haben wir deshalb versucht, geeignete Vokabulare aus der allgemeinen strukturierten Wissensbasis Wikidata abzuleiten. Wikidata ist ein kollaboratives Projekt mit dem Ziel, Wissen als strukturierte Aussagen (Subjekt-Prädikat-Objekt) zu repräsentieren und zu sammeln. Für den Einsatz in der Politikwissenschaft haben wir uns auf Kriege, Wahlen und coup d'états fokussiert und entsprechende Vokabulare aus Wikidata abgeleitet. Ebenfalls setzten wir auf Stanford Stanza[13], ein Named-Entity-Recognition-Tool, um Personen, Organisationen, Orte und Zeitangaben in Texten auszuzeichnen. Die Detektion war dabei jedoch unzuverlässig: Die Vokabulare aus Wikidata enthielten zum Teil unvollständige und vor allem homonyme Begriffe, sodass viele Begriffe aus den Texten falsch aufgelöst wurden. Ebenfalls war die reine Auszeichnung von Personen in Texten, ohne diese zu eindeutigen Bezeichnern aufzulösen, wenig hilfreich. So wurde z.B. Einstein und Albert Einstein jeweils korrekt als Person identifiziert. Jedoch ordnet eine reine Named Entity Recognition diese Namen nicht eindeutig derselben Person zu. Diese problematische Konzeptdetektion führte dann zu Problemen in der Filterung von extrahierten Aussagen. Die extrahierten Aussagen waren zum Teil korrekt, jedoch wenig hilfreich, da diese nicht in Gestalt präziser Konzepte kanonisiert werden konnten. Ebenfalls wurde beobachtet, dass Nominalphrasen als Subjekt einer Aussage oft prägnant und kurz waren, während die Objekte teilweise große Teile eines Satzes umfassten (z.B. einen ganzen Nebensatz). Unsere Ergebnisse sind in unserer vorangegangenen Arbeit detailliert beschrieben[14] und die untersuchten Daten stehen ebenfalls online frei zur Verfügung[15].

Wir zogen deshalb das Fazit, dass eine Übertragung unserer in der Pharmazie eingesetzten Extraktionsverfahren auf die Politikwissenschaft so nicht ohne erheblichen Aufwand möglich ist: Geeignete Konzeptvokabulare und zuverlässige Erkennungsmethoden fehlen und eine Kanonisierung von extrahierten Aussagen bleibt damit erschwert bzw. offen. Wir entschlossen uns daher für ein anderes Vorgehen als in der Pharmazie, d.h. gegen eine strukturierte Repräsentation politikwissenschaftlicher Texte.

13 Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason; Manning, Christopher D.: Stanza. A Python Natural Language Processing Toolkit for Many Human Languages, in: Association for Computational Linguistics (ACL): System Demonstrations, 2020. Online: <https://dx.doi.org/10.18653/v1/2020.acl-demos.14>.

14 Kroll u.a.: Detailed Library Perspective.

15 Toolbox: <https://github.com/HermannKroll/KGExtractionToolbox>, Stand: 05.07.2023.

4. Narrativer Informationszugriff in der Politikwissenschaft

Ein narrativer Informationszugriff im Feld der Politikwissenschaft steht vor mehreren Problemen. Im Unterschied zu den leichter beschreibbaren Verhältnissen zwischen Patient*in, Erkrankung und Wirkstoff sieht man sich bei einschlägigen Textkorpora deutlich komplexeren Aussagen gegenüber, die sich aus umstrittenen und mehrdeutigen Termini mitsamt üppigen Synonymvarianten, variierenden ideologischen Kontexten und insgesamt deutlich längeren Sätzen zusammensetzen. Häufig kehren schon Verneinungen den Sinn komplexer Aussagen um. Entscheidend für die Verarbeitung valider und kohärenter Aussagen ist ein entsprechendes Kontextwissen. Als unmittelbares praktisches Problem ergibt sich damit die Frage: Wie können narrative Suchen in anspruchsvollen Textkorpora überhaupt so funktionieren, dass komplexe Informationszusammenhänge zuverlässig berücksichtigt werden?

4.1 Implementierung des narrativen Zugriffs

Unser Vorhaben fokussierte sich auf die politikwissenschaftliche Subdisziplin[16] der Internationalen Beziehungen[17] bzw. der Weltpolitikforschung[18], die sich mit weltpolitischen Vorgängen, zwischenstaatlichem Handeln und inter- bzw. transnationalen Akteuren und Organisationen befasst. Als Textkorpus diente eine Sammlung zeitgenössischer EU-Parlamentsreden (europarl) aus dem Kontext des Irakkriegs 2003–2011 und die Wikipedia (allgemein bekanntes Kontextwissen). Der eigentliche narrative Informationszugriff wurde wie folgt implementiert. Nutzende formulieren ihre Anfrage als natürlichsprachlichen Satz (Query Statement) sowie mit einem Boole'schen Schlüsselwort-basierten Kontextfilter. Zusätzlich kann die Suche auf europarl bzw. Wikipedia beschränkt und ein minimaler Konfidenzschwellwert für den Vergleich von gesuchtem Satz und Text festgelegt werden. Die Menge relevanter Schlüsselwörter wird mit AND- und OR-Verknüpfungen verbunden, die den Kontext beschränken bzw. erweitern.

Die Implementierung unserer Suche ähnelt einer semantischen Suchanfrage: Die verwendeten Textkollektionen werden in einzelne Sätze aufgeteilt und vorindiziert. Die eingegebenen Schlüsselworte werden verwendet, um relevante Sätze in der ersten Phase zu bestimmen. Dann werden diese Sätze und die eingegebene natürlichsprachliche Aussage mittels eines modernen Sprachmodells[19] verglichen. Die eingesetzte Methode heißt Textual Entailment: Gegeben sind eine Hypothese (z.B. die Aussage/der gesuchte Satz) und eine Prämisse (z.B. ein Satz aus der Kollektion), auf deren Grundlage ein Sprachmodell entscheiden soll, ob die Prämisse die Hypothese unterstützt (entailed), widerlegt (contradicted) oder kein Bezug besteht (neutral). Mit Hilfe dieser Methoden können semantische

16 Die politikwissenschaftlichen Teildisziplinen lassen sich auf der Grundlage aktueller universitärer Kurrikula und einschlägiger Einführungsbände in etwa wie folgt systematisieren: Vergleichende Analyse politischer Systeme bzw. Vergleichende Regierungslehre, Policyforschung, Internationale Politik bzw. Internationale Beziehungen bzw. Weltpolitik, Politische Theorie bzw. Politische Ideengeschichte bzw. Demokratietheorie, Politische Ökonomie, Parteien- und politische Kulturforschung, politische Bildung.

17 Masala, Carlo; Sauer, Frank (Hg.): Handbuch Internationale Beziehungen. Living Reference Work, continuously updated edition, Wiesbaden 2016.

18 Franke, Ulrich; Roos, Ulrich: Rekonstruktive Methoden der Weltpolitikforschung. Anwendungsbeispiele und Entwicklungstendenzen, Baden-Baden 2013.

19 Das eingesetzt Modell (ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli) stammt von Huggingface (einer Open Source Plattform für Sprachmodelle).

Suchen realisiert werden, die auch Re-Phrasierungen oder eine andere Wortwahl auffindbar machen. Jedoch ist der Einsatz von Sprachmodellen kostenintensiv in der eigentlichen Suche. Deswegen wurde für unser Vorgehen die Schlüsselwort-basierte Suche als initialer Schritt eingesetzt, um die auszuwertenden Sätze für das Sprachmodell stark zu reduzieren. Die Textkorpora lagen auf Englisch vor. Das System wurde 2022 implementiert und evaluiert.

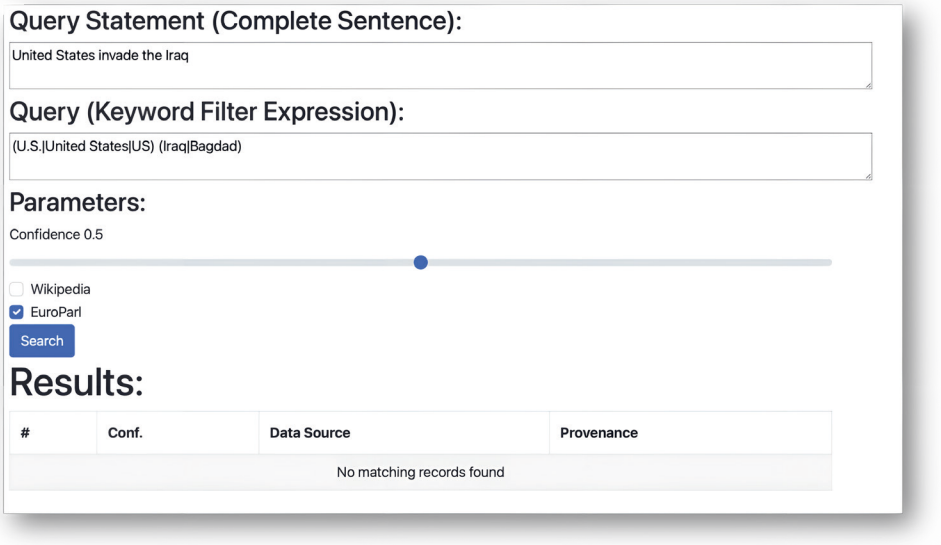

Abbildung 3: Prototypisches Interface der narrativen Suchlösung für die Politikwissenschaft.

4.2 Evaluierung

Das ursprüngliche Ziel der intellektuellen Evaluation war die quantitative Prüfung einer größeren Menge von Suchergebnissen mithilfe eines einfachen binären Kriteriums – „plausibel / nicht plausibel" –, um die Zuverlässigkeit des narrativen Suchsystems im Vergleich mit einer klassischen Schlüsselwort-basierten Suche bewerten zu können. Gleichwohl traten bei ersten Umsetzungsversuchen technische Probleme auf, die sich vorerst nicht lösen ließen. Wir zogen den Schluss, dass die Entwicklung narrativer Suchsysteme noch nicht weit genug vorangeschritten ist, um eine quantitative Analyse zu leisten, und verlagerten den Fokus auf eine stärker explorative qualitative Evaluation der Frage, ob und inwieweit das Suchsystem in der Lage ist, plausible narrative Suchergebnisse zu erzielen. Zu diesem Zweck wurden verschiedene Beispielnarrationen konstruiert, die Argumentationsgänge des europarl-Korpus intellektuell nachbilden, und als Aussagen in das Testsystem eingegeben. Von den 17 z.T. sprachlich variierten Narrativen förderten sechs Aussagen plausible Ergebnisse zutage, die hier auszugsweise wiedergegeben werden sollen:

Aussage I: United States invade the Iraq

Kontext-Filter: *(U.S.|United States|US) (Iraq|Bagdad|Baghdad),*

Beispiel-Ergebnis, Conf. 1.00: „When we were unable to prevent the United States' war in Iraq, or, worse, when we supported it, as most European governments did, we knew of the heightened security risks we were causing in that region, just as we also knew that the next stop after Baghdad might be Damascus."

Abbildung 4: Beispielsuche I und Ergebnisse des entwickelten Suchsystems.

Die zutage geförderten Ergebnisse sind in Bezug auf die eingegebene Aussage durchweg inhaltlich korrekt. Allerdings erzielte das Testsystem in vielen Fällen deutlich häufigere Treffer im besser erschlossenen Wikipedia-Korpus. Während bspw. die Suchanfrage „United States invade the Iraq / (U.S.|United States|US) (Iraq|Bagdad|Baghdad)[20]" mit einem Konfidenzwert von 0,5 nur zwei Treffer im europarl-Korpus liefert, sind es im Wikipedia-Korpus 280 Treffer. Überdeutlich werden hier a) gewisse Anforderungen an die Phrasierungs- bzw. Formulierungskompetenz der Nutzenden, da sämtliche Suchbegriffe zwingend in Kombination vorgefunden werden müssen, und b) die Vorzüge – oder auch: das Erfordernis – einer qualitativ hochwertigen Erschließung der zugrunde liegenden Daten bzw. Texte, da Mehrdeutigkeiten und Begriffsvarianten im Kontext politischer Argumentationen eine große Rolle spielen.

Die mit Abstand zufriedenstellendsten Ergebnisse liefert hier Suchanfrage VI:

Aussage VI: **European Union should act to secure peace**

Kontext-Filter: *(European Union | EU)*

Ergebnisse (Auswahl aus insgesamt 20 Resultaten):

Conf. 1.0: „Latvia wholeheartedly advocates the assumption of greater responsibility by the EU in ensuring worldwide peace and security, and believes that the new European security strategy is a step in the right direction."

Conf. 0.96: „After this motion has been approved, the European Union shall propose, through its representatives on the Security Council, that a UN military force be sent to keep the peace between the Israeli and Palestinian states."

Conf. 0.91: „The foreign policy of the new and expanded Union must therefore endeavour to enhance the security and promote the prosperity of the EU's border regions."

Conf. 0.88: „In the statements made today by representatives of the European Union, I did not detect such strong commitment to ensure that the European Union actually does all that lies within its power to stop the widespread slaughter in Sudan."

Conf. 0.54: „Wherever terrorism lurks, wherever it rears its head, whatever form it takes, the European Union must not shrink from fighting it."

Abbildung 5: Beispielsuche VI und Ergebnisse des entwickelten Suchsystems.

20 Die Notation bedeutet: (U.S. OR United States OR US) AND (Iraq OR Bagdag OR Baghdad)

Nicht nur fällt die Trefferzahl mit n=20 vergleichsweise üppig aus, auch gelingt es dem Suchsystem, ein Argument mit vergleichsweise hohem Abstraktionsniveau in verschiedenen Ausprägungen der Dokumentenkollektion zu identifizieren bzw. auffindbar zu machen. Oder anders formuliert: In dieser Suchanfrage wird eine für eine Proto-Diskursanalyse denkbare Forschungshypothese gewissermaßen praktisch suchbar: Dass die Europäische Union (EU) zum Ziel der Friedenssicherung aktiv werden soll, wird als allen gesuchten Textausschnitten gemeinsame Forderung in der Suchphrase eingegeben. Aus den Ergebnissen schließlich geht die jeweilige argumentative Struktur zur Begründung der Frage, wie die EU dieses Ziel erreichen soll, hervor – eine Struktur, deren Passung zu einem klassischen weltpolitischen Argument nun untersucht werden kann. Dabei demonstriert besonders das zuletzt aufgeführte Ergebnis mit dem verhältnismäßig niedrigen Konfidenzwert eine politisch hochaufgeladene Aussage, welche zum eingegebenen narrativen Muster passt und so nicht von einer klassischen Schlüsselwort-basierten Suche hätte gefunden werden können.

4.3 Diskussion

Auf der einen Seite zeichnen die Testresultate ein ambivalentes Bild: Trotz des erzielten Ertrags steckt insbesondere die Möglichkeit zur Paraphrasierung von Informationen in Satzstrukturen höheren Abstraktionsniveaus zum jetzigen Zeitpunkt noch in ihren Anfängen (bspw. mit Subjekten wie „Akteur“ oder „Staat“ statt „United States“ in den oben aufgeführten Beispielnarrationen). Damit ließ sich der Nutzen von größeren, komplexeren narrativen Mustern noch nicht vollständig prüfen. Eine erneute, technisch deutlich aufwändigere Weiterentwicklung war aufgrund der zeitlichen und personellen Limitationen vorerst nicht umsetzbar. Auf der anderen Seite beweisen diese ersten Ergebnisse, dass narrative Suchlösungen, die mit Primärtext-Korpora arbeiten, für die politikwissenschaftliche Forschung und Lehre nutzbar gemacht und hilfreich aufbereitet werden können. Denn Anfragen mit einfachen Narrationen sind prinzipiell schon jetzt möglich: Sie können im Sinne eines „Trial and Error“-Verfahrens einer ersten Prüfung von Hypothesen für diskurs- und inhaltsanalytische Forschungsvorhaben dienen. Künftig könnte ein fortgeschritteneres System in der Lage sein, Suchanfragen höheren Abstraktionsgrades und größerer thematischer Breite zu bedienen. So wäre es möglich, Teile eines intellektuellen Verfahrens im Forschungsprozess vorbereitend durch ein maschinelles zu ersetzen und damit die Orientierung in großen Textmengen, perspektivisch auch – und das ist entscheidend – ohne Vorkenntnisse der zugrunde gelegten Dokumentenkollektion, maßgeblich zu erleichtern. Damit solche komfortablen, forschungsunterstützenden Services praxistauglich funktionieren, müssen passende Plausibilitätskriterien implementiert, eine hinreichende Datenbasis erschlossen sowie eine qualifizierte Phrasierung[21] der als Narrationen formulierten Suchanfragen unterstützt werden.

4.4 Ausblick

Die aufgeführten aktuellen Defizite wie die (noch) nicht anwendbaren Extraktionsverfahren sowie Limitierungen des implementierten Suchsystems sind keineswegs mit einem Scheitern des Gesamtvorhabens gleichzusetzen. Im Gegenteil: Die erzielten konzeptuellen und explorativ-empirischen Fortschritte des narrativen Informationszugriffs zur Unterstützung der politikwissenschaftlichen

21 Eines der großen Hindernisse liegt nach wie vor im geringen Erschließungsgrad nutzbarer Narrationen.

Forschung legen vielversprechende Grundsteine. Die fortschreitende Entwicklung von Sprachmodellen wird, wie z.B. ChatGPT zeigt, in der Zukunft narrative Suchlösungen weiter unterstützen können. Dabei sollte es nicht das Ziel sein, eine Antwort ohne entsprechende Quellenverweise/Belege zu erzeugen. Unser Meinung nach sollten Anfragen weiterhin in validen Kontexten beantwortet werden. In unserem Szenario bedeutet das die Beantwortung mit geeigneter Literatur bzw. einschlägigen Primärtexten, die hinsichtlich ihres Kontexts kompatibel sind[22].

5. Fazit

Ein narrativer Informationszugriff, wie er derzeit in PubPharm angeboten wird, ist grundsätzlich auf politikwissenschaftliche Forschungskontexte übertragbar. Allerdings muss aufgrund fehlender Konzeptvokabulare sowie geeigneter Extraktionsverfahren ein anderer Ansatz als in der Pharmazie gewählt werden: Die Konvertierung von Texten in strukturierte Repräsentationen ist in der Politikwissenschaft erheblich aufwändiger und so nicht in absehbarer Zeit möglich. Stattdessen stellt der hier vorgestellte Prototyp den ersten Schritt in Richtung eines Suchsystems dar, mit dessen Hilfe politikwissenschaftliche Forschungshypothesen als Anfragen in Textkorpora aus Quellen und Primärtexten eingespeist werden können. Anstatt einer Vorverarbeitung der Texte wird auf einer semantischen Suche aufgebaut, ermöglicht durch moderne Sprachmodelle. Bei einer künftigen Weiterentwicklung des hier getesteten Systems käme es in erster Linie darauf an, Sätze mit höheren Abstraktionsniveaus suchbar zu machen, damit narrative semantische Muster im Sinne von vorbereitenden Diskurs- und Deutungsmusteranalysen umfänglich eingegeben werden können.[23] Hier wäre dann insbesondere auch ein technischer Weg wünschenswert, wie man größere, komplexere Muster eingeben kann, um der semantischen Mehrteiligkeit von politischen Deutungsmustern gerecht zu werden.

Literaturverzeichnis

- Banko, Michele; Cafarella, Michael J.; Soderland, Stephen; Broadhead, Matthew; Etzioni, Oren: Open Information Extraction from the Web, in: IJCAI, 2007. Online: <http://ijcai.org/Proceedings/07/Papers/429.pdf>.
- Draheim, Christina; Keßler, Kristof; Wawrzinek, Janus; Wulle, Stefan: Die Rechercheplattform PubPharm, in: GMS Medizin – Bibliothek – Information 19 (3), 2019. Online: <https://dx.doi.org/10.3205/mbi000448>.
- Franke, Ulrich; Roos, Ulrich: Rekonstruktive Methoden der Weltpolitikforschung. Anwendungsbeispiele und Entwicklungstendenzen, Baden-Baden 2013.
- Keßler, Kristof; Kroll, Hermann; Wawrzinek, Janus; Draheim, Christina; Wulle, Stefan; Stump, Katrin; Balke, Wolf-Tilo: PubPharm – Gemeinsam von der informationswissenschaftlichen Grundlagenforschung zum nachhaltigen Service, in: ABI Technik 39 (4), 2019. Online: <https://doi.org/10.1515/abitech-2019-4005>.

22 Im Beispiel der Pharmazie könnte dieser Kontext durch ähnliche Studienbedingungen abgebildet werden. In der Politikwissenschaft sind gleiche Weltanschauungen, Standpunkte oder Denkschulen geeignete Beispiele.

23 Ein weiterer Nutzen des Systems könnte in der Suggestion möglicher paralleler Forschungshypothesen am Gegenstand des jeweiligen Dokumentenkollektionen liegen. Ähnliches leistet bereits PubPharms narrativer Service, hier allerdings für wissenschaftliche Literatur, die mit einem einheitlichen Fachvokabular arbeiten kann.

- Kroll, Hermann; Kalo, Jan-Christoph; Nagel, Denis; Mennicke, Stephan; Balke, Wolf-Tilo: Context-Compatible Information Fusion for Scientific Knowledge Graphs, in: International Conference on Theory and Practice of Digital Libraries, 2020. Online: <https://doi.org/10.1007/978-3-030-54956-5_3>.
- Kroll, Hermann; Draheim, Christina: Narrative Information Access for a Precise and Structured Literature Search, in: o-bib 8 (4), 2021, S. 1–13, <https://doi.org/10.5282/o-bib/5730>.
- Kroll, Hermann; Pirklbauer, Jan; Kalo, Jan-Christoph; Ruthmann, Johannes; Balke, Wolf-Tilo: A Discovery System for Narrative Query Graphs. Entity-Interaction-Aware Document Retrieval, in: International Journal on Digital Libraries, 2023. Online: <https://doi.org/10.1007/s00799-023-00356-3>.
- Kroll, Hermann; Pirklbauer, Jan; Balke, Wolf-Tilo: A Toolbox for the Nearly-Unsupervised Construction of Digital Library Knowledge Graphs, in: ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2021, S. 21–30. Online: <https://doi.org/10.1109/JCDL52503.2021.00014>.
- Kroll, Hermann; Pirklbauer, Jan; Plötzky, Florian; Balke, Wolf-Tilo: A Detailed Library Perspective on Nearly Unsupervised Information Extraction Workflows in Digital Libraries, in: International Journal on Digital Libraries, 2023. Online: <https://doi.org/10.1007/s00799-023-00368-z>.
- Masala, Carlo; Sauer, Frank (Hg.): Handbuch Internationale Beziehungen. Living Reference Work, continuously updated edition, Wiesbaden 2016.
- Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason; Manning, Christopher D.: Stanza. A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL): System Demonstrations, 2020. Online: <http://dx.doi.org/10.18653/v1/2020.acl-demos.14>.
- Thaller, Manfred: Controversies Around the Digital Humanities. An Agenda, in: Historical Social Research 37 (3), 2012, S. 7–23. Online: <https://doi.org/10.12759/hsr.37.2012.3.7-23>.
- Vrandečić, Denny; Krötzsch, Markus: Wikidata. A Free Collaborative Knowledgebase, in: Communications of the ACM 57 (10), 2014, S. 78–85. Online: <https://dx.doi.org/10.1145/2629489>.