

Von A bis Z: Überlegungen zur Erstellung eines Wissensgraphen aus historischen Enzyklopädien

Hagen, Thora

thora.hagen@uni-wuerzburg.de
Universität Würzburg, Deutschland

Strukturierte Daten, insbesondere Wissensgraphen, gewinnen in vielen Forschungsfeldern zunehmend an Bedeutung. Sie bieten eine kondensierte Sicht auf menschliches Wissen, sei es allgemein oder domänen-spezifisch. Unter anderem können sie damit in Forschungsprojekten helfen, Textkorpora mit zusätzlichen Informationen anzureichern oder das Beantworten spezifischer Fragestellungen durch Inferenzbildungen erst möglich zu machen. Für den Erstellungsprozess eines Wissensgraphen gibt es allerdings keinen universal gültigen Lösungsweg (Kejriwal 2019, 4).

In diesem Beitrag geht es speziell um die Herausforderungen, einen Wissensgraphen aus historischen Enzyklopädien zu erstellen. Als Beispiel dafür dient das DFG-geförderte Projekt EncycNet.¹ Es soll hierbei nicht die Präsentation von Methoden und deren Ergebnisse im Vordergrund stehen, sondern es sollen eher die Besonderheiten des Vorhabens und die damit einhergehenden konzeptuellen Überlegungen, die letztendlich auch bei der Auswahl der Methoden eine Rolle spielen, genauer dargestellt werden. Im Folgenden soll deshalb zunächst ein kurzer Überblick über Wissensgraphen und deren Erstellung gegeben werden. Den Hauptteil leitet dann eine kurze Einführung zu EncycNet ein und schließlich werden die drei Herausforderungen des Projekts diskutiert.

Forschungskontext

Um zu verstehen, was die Umwandlung von verschiedenen Quellen in einen Wissensgraph bedeutet, sollte zunächst der Begriff „Wissensgraph“ geklärt werden. Es gibt zwei verschiedene Definitionen des Begriffs; eine traditionelle und eine moderne Verwendung. Traditionell wird davon ausgegangen, dass ein Wissensgraph nur Weltwissen enthält – ganz konkret nur Named Entities. Dabei wird auch erwartet, dass es eine schwergewichtige Ontologie zu dem Wissensgraph gibt. Geprägt ist diese traditionelle Sichtweise durch das erste Aufkommen des Begriffs durch den Google Knowledge Graph, also Wikidata, welcher genau diese Eigenschaften besitzt. Gemäß dieser Definition sind also WordNet oder GermaNet auch keine Wissensgraphen, sondern semantische Netzwerke. Die moderne Definition ist dem

gegenüber etwas weniger streng; hier können Wissensgraphen jegliche Art von Wissen abbilden und die Ontologie darf ebenso auch leichtgewichtig sein.

Die Erstellung von Wissensgraphen lässt sich in drei Schritten zusammenfassen: 1) der Aufbau einer Ontologie, 2) die Erkennung von Entitäten und deren Alignierung und 3) die Relationsextraktion. Zusammengefasst aus der neuesten Forschung in dem Feld (aus Chaves-Fraga et al. 2021 u. 2022) lässt sich sagen, dass hier hauptsächlich das Mapping von Datenstrukturen im Vordergrund steht bzw. das Vereinheitlichen verschieden strukturierter Daten, um einen Graphen zu erstellen. Das bedeutet: Für alle drei Schritte kann auf bereits strukturierte Information zurückgegriffen werden und die Re-Organisation ist Kern der Aufgabe (Schröder et al. 2021, Wu et al. 2020). Daneben stehen häufig auch domänenspezifische Anforderungen an den Graphen im Vordergrund, so wie zum Beispiel bei der Erstellung des Open Drug Knowledge Graphs (Mann et al. 2021).

Ein Teilbereich der Forschung wiederum beschäftigt sich auch mit dem Erstellen von Graphen aus Fließtext, wobei hier die traditionelle Sichtweise auf Wissensgraphen, also die Repräsentation von realen Entitäten, dominiert. Typische Methoden aus dem Bereich sind daher z.B. Named Entity Recognition und Entity Linking (Kejriwal 2019, 12, 33). Dies gilt auch für das Feld der Digital Humanities, eben besonders für das Beschreiben kultureller Objekte, so wie zum Beispiel dem Modellieren von Erzählorten in Romanen (Hinzmann et al. 2022) oder dem Modellieren von Künstlern und Werken aus kunsthistorischen Texten (Jain et al. 2022). Daneben können auch syntaktische Marker statt Named Entities die Graphmodellierung stützen (Perak 2020). Häufig wird auch auf strukturierte Daten, z.B. Wissensgraphen aus derselben Domäne, zurückgegriffen, um eine Basis für die Alignierung der Konzepte, Relationsauswahl und Ontologie zu schaffen (Jain et al. 2022, Clancy et al. 2019).

EncycNet: Herausforderungen und Chancen

Das Ziel von EncycNet ist es, einen Wissensgraphen aus sechs historischen Enzyklopädien (genauer: Konversationslexika, siehe Tabelle 1) zu erstellen.² Der Graph soll ein lexikalisch-semantisches Netzwerk sein, welcher einerseits die Einträge über alle Enzyklopädien hinweg aligniert und disambiguiert. Zum anderen müssen die relevanten Relationen aus den Einträgen enthalten sein. So sollen beispielsweise Themenbereiche zu Stichwörtern verzeichnet sein („Operation“ / „Mathematik“ und „Operation“ / „Medizin“), aber auch typische Informationen über Named Entities wie Geburtsdatum und Geburtsort von bekannten Personen. Der Graph soll zunächst eine Vogelperspektive auf die Daten bieten; insbesondere sollen so Bedeutungsverschiebungen eines Konzepts im Verlauf der Zeit (z.B. durch Veränderung der nächsten Nachbarn) sichtbar gemacht werden. Ebenfalls kann mithilfe von Inferenzen der Graph genauer analysiert werden: Etwa die Verwandtschaftsgrade von Konzepten über Pfade, Zentralität von Konzepten für eine bestimmte Zeit oder die Bildung von Communities. Darüber hinaus können die Graphdaten dazu dienen,

Textkorpora anzureichern oder historische Evaluationsdaten bereitzustellen.

Tabelle 1: Auflistung der Konversationslexika für die Grundlage des Wissensgraphen EncycNet

Lexikon	Anzahl Einträge	Anzahl Tokens
Brockhaus Conversations-Lexikon oder kurzgefaßtes Handwörterbuch (1809)	6.960	1.186.000
Brockhaus Bilder-Conversations-Lexikon (1837)	7.049	2.604.000
Brockhaus Kleines Konversations-Lexikon (1911)	82.780	2.434.000
Damen Conversations Lexikon (1834)	7.099	1.461.000
Herders Conversations-Lexikon (1854)	39.755	2.256.000
Meyers Großes Konversations-Lexikon (1905)	156.264	17.437.000

Für die Erstellung des Graphen bedeutet der dargestellte Forschungskontext, dass ein solches Vorhaben verschiedene Forschungsnischen in sich vereint. Zum einen bedient sich EncycNet der modernen Definition für Wissensgraphen. Da in den historischen Enzyklopädien nicht nur Personen und Orte zu finden sind, sondern auch Objekte und Phänomene aller Art, muss der resultierende Wissensgraph diese Dinge auch abbilden können. Zum anderen liegen die Enzyklopädien nur semi-strukturiert (TEI-XML) vor.

³ – zum Großteil handelt es sich deshalb um eine Grapherstellung aus Fließtext. Und letztlich, durch die Diversität, die mit einem historischen Korpus einhergeht, muss die Erstellung zusätzlich auf heterogene Fließtextdaten abgestimmt sein. Auf diese drei Herausforderungen, die das Enzyklopädienkorpus mit sich bringt, soll im Folgenden genauer eingegangen werden.

Domänenübergreifendes Wissen

In den Enzyklopädien wird jegliche Art von Wissen über unterschiedlichste Wissensdomänen äußerst detailliert abgebildet. Insbesondere dann, wenn Einträge deutlich länger als nur eine Definition sind, findet sich dort einiges an Wissen, welches sich idealerweise auch in dem resultierenden Wissensgraphen widerspiegeln sollte. Allerdings sind solche längeren Einträge auch weniger standardisiert als die kürzeren, einfachen Begriffsdefinitionen in den Enzyklopädien. Trotzdem gibt es auch in längeren Einträgen bereits einige vorstrukturierte Elemente, gekennzeichnet durch die XML Annotation, die sich auf den ersten Blick zum extrahieren anbieten. Dazu gehören beispielsweise Hierarchien oder Aufzählungen, Vers, oder auch semi-strukturierte Formen im Fließtext wie z.B. Gleichungen, Angaben von Einheiten oder Ähnliches (einige Beispiele in Abbildung 1). Allerdings sind diese Elemente dann fast immer domänenspezifisch und sind damit nur in den wenigsten Einträgen zu finden.

"Cam" in Herder 1854

Cam, ostind. Silbermünze = 4 Sgr. $9\frac{1}{4}$ Pf. = $15\frac{1}{2}$ kr. C.-M.

"Pint" in Herder 1854

Pint, engl. und nordamerik. Hohlmaß; für Getreide = $28\frac{3}{8}$, für Flüssigkeit = $28\frac{13}{25}$ Par. Kubikzoll. P.e. Getreidemaß in der Lombardel = 50,4 Par. Kubikzoll; Flüssigkeitsmaß in Bergamo = 62,1, in Brescia = $69\frac{1}{2}$, auf

"Apotheke" in Herder 1854

worden sein. – Apothekergewicht in Deutschland: Pfund, As (℔) = 12 Unzen, (3), die U. = 8 Drachmen (3), die D. = 3 Skrupeln (3), der Skr. = 20 Gran (*gr.*), von denen also 5760 auf 1 Pf. gehen.

"Asien" in Meyer 1905

Sprachen gibt für die Bevölkerung Asiens ungefähr folgende Gruppen:

A. Nordasien.

I. Inkgirisch. II. Korjakisch, Tschuktschisch. III. Sprachen von Kamtschakta und Kurilen (Aino). IV. Jenissei-Ostjakisch und Kottisch.

B. Mittel- oder Hochasien.

I. Uralaltaische Sprachen. a) Samojedische Gruppe: Jurakisch,

"Geschwindigkeit" in Brockhaus 1837

In einer Sekunde legen zurück:

Flüsse von mittlerer Geschwindigkeit	3–4 Fuß.
Winde von mäßiger Stärke	10 Fuß.
Ströme von größter Geschwindigkeit	12 Fuß.

"Assonanz" in Brockhaus 1837

nichts Weicheres und Lieblicheres denken, als Verse wie diese:

»Wonne weht von Thal und Hügel,

Weht von Flur und Wiesenplan, Weht vom glatten Wasserspiegel,

Wonne weht mit welchem Flügel

Des Piloten Wange an.«

Abbildung 1: Beispiele für strukturierte bis semi-strukturierte Inhalte in den Einträgen der Enzyklopädien.

Um ein möglichst generisches Bild von dem Inhalt der Enzyklopädien zu erhalten, können die Einträge zunächst in generische Klassen unterteilt werden; für EncycNet ergaben sich die Klassen Personen, Orte, Objekte und Abstrakta. Diese Klassen wurden nach dem Vorbild von Wikidata (Personen und Orte als 2 Hauptkategorien) und WordNet („physical entity“ und „abstract entity“) als direkte Hyponyme des Wurzelements „entity“. Mehrere Artikel aus diesen Klassen können dann gesichtet werden und in Abstimmung mit Wikidata oder WordNet die wichtigsten Informationen oder thematische Segmente innerhalb der Einträge identifiziert werden.

So ergeben sich drei Gütekriterien für die regelbasierte Extraktion von strukturiertem Wissen aus historischen Enzyklopädien: 1) Wie generisch ist die Information; also auf wie viele Klassen und auf wie viele Einträge trifft sie insgesamt zu, 2) Wie stark ist die Information vorstrukturiert, also wie präzise kann eine Regel für die Information gefunden werden und 3) Wie relevant ist die Information im Hinblick auf die Ziele, die der Wissensgraph verfolgt? Recht generische Informationen sind beispielsweise Synonyme und Übersetzungen; in den meisten Einträgen werden diese direkt nach der Nennung des Konzepts aufgelistet und sind damit auch vergleichsweise

einfach zu extrahieren. Die Beispiele aus Abbildung 1 sind das Gegenteil: Zwar sind sie alle eher einfach zu extrahieren durch die vorstrukturierte Form, allerdings sind sie auch in nur wenigen Einträgen vorhanden und insbesondere Zahlen und Einheiten sind für das Ziel das EncycNet verfolgt, nämlich ein semantisches Netzwerk zu bilden, eher uninteressant. Es empfiehlt sich daher, zunächst alle möglichen Informationen nach diesen Kriterien zu sortieren und dann erst mit der Extraktion zu beginnen, um eben nicht nur Detailwissen zu extrahieren bzw. bestimmte Domänen zu bevorzugen.

Die domänenübergreifende Perspektive wirkt sich auch auf die Auswahl der Ontologie aus. Zusammen mit dem Ziel, nicht nur Entitäten sondern auch lexikalischen Wissen semantisch zu modellieren, schließt dies einige Standards aus. OntoLex (Cimiano et al. 2016) beispielsweise ist gerade dafür gedacht, lexikalisches Wissen aus Nachschlagewerken in einen Graphen umzuwandeln. Allerdings liegt hier der Fokus auf morphologischen statt semantischen Eigenschaften (Wortart, Genus, etc.), welche in Lexika nicht unbedingt aufgeführt werden. Die Struktur des Lexikons wird außerdem explizit beibehalten. So werden beispielsweise Referenzen auf andere Artikel als Kante „reference“ eingepflegt und nicht weiter typisiert. Daneben gibt es CIDOC-CRM (Doerr 2005), eine Standard-Ontologie zum Beschreiben von kulturellen Objekten, und Faktoide (Bradley und Short 2005) zur Abbildung von spezifischen Stellen in einer Quelle über Personen. Bei beiden Modellierungsarten sind Objekte und Eigenschaften eher auf Named Entities und weniger auf Lexeme ausgelegt. Alle drei Standards sind gut geeignet für Teilbereiche der Enzyklopädien, so wie zum Beispiel Faktoide für biographische Einträge. Die möglichst vollständige Abbildung aller Inhalte die EncycNet anstrebt, auch hauptsächlich von lexikalischen Eigenschaften (z.B. Synonyme oder Hyperonyme), ist aber nicht möglich.

Synonyme und Hyperonyme gehören bei der Bildung eines lexikalisch-semantischen Netzwerks zu den Grundbausteinen; sind also besonders wichtig für Punkt 3. Synonyme werden benötigt, um Synsets nach dem Vorbild von WordNet zu bilden (bzw. das Verknüpfen gleicher Konzepte) und Hyperonyme für den Aufbau einer Taxonomie. Allerdings ist die Extraktion einer vollständigen Taxonomie, welche alle Domänen umfasst, nur aus dem Enzyklopädienkörper weitestgehend unrealistisch, weswegen auf zusätzliches Material zurückgegriffen werden muss. Als bislang größte Online-Enzyklopädie kann Wikidata / Wikipedia, welche inzwischen nicht nur Entitäten-bezogenes Wissen sondern über die Integration von WordNet auch über lexikalisches Wissen verfügt, als Schnittstelle genutzt werden. Sie liefert einerseits die Taxonomie, aber auch andererseits über Wikipedia zusätzliches Textmaterial zu den Konzepten, welches ebenfalls für die Alignierung der Einträge über verschiedene Enzyklopädien hinweg von nutzen sein kann (Hagen et al. 2022).

Heterogene Daten

Der Wissensgraph, der durch EncycNet entstehen soll, fasst das Wissen von 6 allgemeinen Enzyklopädien zusammen. Alle diese Enzyklopädien unterscheiden sich

hinsichtlich der Auswahl und Organisation der Konzepte, der inneren Struktur der Einträge, Umfang und Auslegung der Definitionen und dem Stil. Jede dieser Eigenheiten müssen für die Informationsextraktion berücksichtigt werden, was bedeutet, dass die Relationen für alle Enzyklopädien neu beurteilt werden müssen. Pragmatisch betrachtet heißt dies auch, dass die Informationsdichte im resultierenden Graph abnimmt, da weniger Relationen in Betracht genommen werden können. Insbesondere auf mögliche Inferenzbildungen durch den Graph wirkt sich das negativ aus.

Aus diesem Grund sollten generische Methoden für die Informationsextraktion hinzugezogen werden, so dass der Graph an Informationsdichte gewinnt. Hierbei kann auf typische Methoden in den Digital Humanities zurückgegriffen werden: Topic Modeling zum Identifizieren von übergreifenden Wissensbereichen, TF-IDF für distinktive Terme eines Eintrags, Komposita des Konzepts als verwandte Konzepte, Named Entity Recognition, oder spezifisch für Enzyklopädien die Extraktion von Referenzen auf andere Einträge. Nachteil dieser generischen Extraktion ist allerdings, dass das Mapping der extrahierten Terme auf eine Relation sich schwieriger gestaltet bzw. viele der Terme eine unspezifische Relation zum Konzept erhalten (z.B. in GermaNet „related_to“).

Auch für die Alignierung ergeben sich praktische Probleme bei der Grapherstellung, denn in den Enzyklopädien können die Konzepte unterschiedlich organisiert sein. Dies reicht von Betitelungskonventionen der Einträge (z.B. „Der Adler“ / „Adler“, oder „William Shakespeare“ / „Shakespeare“ / „Shakespeare, William“) bis hin zu der Möglichkeit, dass es Einträge gibt, die mehrere Konzepte zusammenfassen. In Meyer (1905) werden beispielsweise manche Entitäten thematisch gruppiert. So gibt es zum Beispiel genau zwei Einträge zu *Alexander*: einer gruppiert Fürsten und der andere griechische Schriftsteller mit dem Vornamen, die auch jeweils nochmal genauer beschrieben werden. In Herder (1854) dagegen gibt es drei Einträge: einen, der den Vornamen ohne Personenzuordnung nennt, einen zu Alexander I. (welcher auch Alexander II. umfasst) und einen zu Alexander III. Hier gilt es also zu klären, inwieweit die Konzepte getrennt werden können. In Meyer sind die Konzepte deutlich durch Paragraphen gekennzeichnet, in Herder werden sie sprachlich vermischt. Zusätzlich können allerdings, selbst wenn die Konzepte isoliert sind, diese auch unterschiedlich ausgelegt werden, auch historisch bedingt.

Historisches Wissen

Die Historizität der Daten wirkt sich damit ebenso auf die Alignierung der Konzepte aus. Auf der einen Seite werden teils archaische Begriffe oder Schreibweisen für Konzepte verwendet (z.B. „Irrenanstalt“, jetzt „psychiatrische Klinik“). Durch eine orthographische Normalisierung und durch die Verwendung von Wikipedia für die Alignierung, welche zum Teil auch archaische Begriffe umfasst und auflösen kann, kann diesem Problem entgegengekommen werden. Auf der anderen Seite können sich aber auch die Definitionen von Begriffen im Laufe der Zeit stark verändert haben, etwa weil sich ein Konzept

in ein anderes verwandelt hat oder weil Konzepte zu unterschiedlichen Zeiten anders ausgelegt werden.

So wird beispielsweise in Meyer (1905) die *Exploration* beschrieben als die physische Untersuchung eines Kranken durch einen Arzt, während sie in Wikipedia mit *Anamnese* gleichgestellt wird, also der Erfragung von Informationen im Rahmen einer Erkrankung. Ein weiteres Beispiel findet sich in Herder (1854): Der Begriff *Proportionalität* wird beschrieben durch „Harmonie der Größenverhältnisse“ und „Proportionslehre der menschlichen Gestalt,“ während für *Proportion* die Verhältnisgleichung in der Mathematik genannt wird; also was die Menschen heute eigentlich unter Proportionalität verstehen würden.

Diese Beispiele zeigen eine notwendige Modellierungsentscheidung für einen historischen Wissensgraphen auf: Sollen Konzepte, welche sich grundlegend verändert haben trotzdem miteinander aligniert werden und die Alignierung macht somit den semantischen Wandel sichtbar? Oder sollten nur Konzepte, welche tatsächlich semantische Äquivalente sind aligniert werden, da die Definitionen so unterschiedlich ausfallen können? Je nach Ziel, den der resultierende Graph verfolgt, kann diese Entscheidung unterschiedlich ausfallen.

Letztlich wirkt sich das historische Wissen auch auf den Aufbau der Taxonomie aus. Um eine vollständige Taxonomie automatisch zu generieren, kann auf bestehendes strukturiertes Wissen (wie Wikidata) zurückgegriffen werden. Allerdings besteht hierbei die Gefahr, historisches Wissen zu ignorieren oder zu überschreiben. In den Enzyklopädien wird beispielsweise der Begriff *Hexe* als „Unholdin,“ „Weib“ oder „weissagende Frau“ beschrieben, während in Wikidata „Magier“ verwendet wird. Ein anderes Beispiel ist *Hierarchie*, welche in Wikidata als „Struktur“ oder „System“ eingeordnet wird, jedoch in den Enzyklopädien mit „Priesterherrschaft,“ „Macht“ oder „alle Rechte der Römischen Päpste über die gesamte Christenheit“ beschrieben wird.

Zusammenfassung

Das Bilden eines historischen Wissensgraphen aus Enzyklopädien kombiniert zwei Forschungsnischen aus dem Forschungsfeld. Einerseits geht es hier um die Abbildung von lexikalisch-semantischem Wissen und nicht um nur um Entitätenwissen wie z.B. Wikidata, und andererseits stellt heterogener Fließtext die Datengrundlage für den Graphen dar. Beides sind Voraussetzungen, die sich selten in der aktuellen Forschung zur Erstellung von Wissensgraphen wiederfinden. Der Aufbau von EncycNet wird geleitet von praktischen Anforderungen an den Graphen, welche sich aus den vordefinierten Zielen ergeben. Dabei geht es um die Einschätzung, welche Informationen in den Enzyklopädien sich für eine Relationsextraktion anbieten, aber gleichzeitig soll möglichst viel Wissen mit dem Graphen abgedeckt werden. Zusätzlich, insbesondere um Inferenzen zu ermöglichen, muss der resultierende Graph eine möglichst große Informationsdichte in Tiefe (Taxonomie) und Breite (generische Relationen) aufweisen. Es ergibt sich daraus eine Bottom-up Strategie (schematisch zusammengefasst in Abbildung 2).

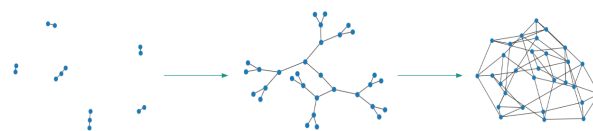


Abbildung 2: Schematische Darstellung der Erstellung des Graphen: gezielte Extraktion von Relationen aus den Enzyklopädien (links), Aufbau der Taxonomie (Mitte) und Verdichtung durch ungezielte, generische Relationsextraktion (rechts).

Für EncycNet werden aktuell Alignierung sowie Relationsextraktion fertiggestellt. Da die größtmögliche Abdeckung für beide Aufgaben erzielt werden soll, werden die Methoden diesbezüglich kontinuierlich optimiert. Noch ausstehend ist die Evaluierung des extrahierten Wissens mit Golddaten. Final sollen dann über die Evaluierung die Relationen und die Alignierung mit Gewichten ausgestattet werden, welche die Konfidenz angeben. Im Frühjahr 2024 sollen dann die Daten in RDF* zur Verfügung gestellt werden.

Im Vordergrund dieses Beitrags sollten damit jene Entscheidungsfindungen stehen, die in Methoden-orientierten Beiträgen sonst meist nur am Rande erwähnt werden. Dabei wurde sich auf EncycNet bezogen, jedoch können die hier aufgeführten Ideen genauso für Projekte, die auf ähnliche Herausforderungen bei dem Aufbau eines Wissensgraphen stoßen, interessant sein.

Fußnoten

1. <https://encycnet.github.io/>
2. Eine ausführliche Übersicht über das Korpus ist auf <https://encycnet.github.io/corpus-overview.html> gegeben.
3. Die TEI-Daten können über Zenodo (<http://dx.doi.org/10.5281/zenodo.4039569>) heruntergeladen werden.

Bibliographie

- Bradley, John und Harold Short.** 2005. "Texts into databases: the Evolving Field of New-style Prosopography." *Literary and Linguistic Computing* 20 (1): 3-24.
- Chaves-Fraga, David, Anastasia Dimou, Pieter Heyvaert, Freddy Priyatna und Juan Sequeda.** Hrsg. 2021. „Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021).“ In *CEUR Workshop Proceedings* 2873. <http://ceur-ws.org/Vol-2873/> (zugegriffen: 01. August 2022).
- Chaves-Fraga, David, Anastasia Dimou, Pieter Heyvaert, Freddy Priyatna und Juan Sequeda.** Hrsg. 2022. „Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022).“ In *CEUR Workshop Proceedings* 3141. <http://ceur-ws.org/Vol-3141/> (zugegriffen: 01. August 2022).
- Cimiano, Philipp, John P. McCrae, und Paul Buitelaar.** 2016. *Lexicon model for ontologies: community report*, 10

May 2016. <https://www.w3.org/2016/05/ontolex/> (zugegriffen: 07. Dezember 2022).

Clancy, Ryan, Ihab F. Ilyas und Jimmy Lin. 2019. „Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond.“ In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, Hong Kong, China.

Doerr, Martin. 2005. "The CIDOC CRM, an ontological approach to schema heterogeneity." *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Hagen, Thora, Fotis Jannidis und Andreas Witt. 2022. „Word sense alignment and disambiguation for historical encyclopedias.“ In *6th International Conference on Graphs and Networks in the Humanities*. urn:nbn:de:bsz:mh39-109834 (zugegriffen: 09. Dezember 2022). Vorveröffentlichung.

Hinzmann, Maria, Julia Röttgermann, Anne Klee, Moritz Steffes und Christof Schöch. 2022. „The French Enlightenment Novel as a Graph? Potentials and Challenges in the Construction of a Knowledge Network.“ In *6th International Conference on Graphs and Networks in the Humanities*. 10.5281/zenodo.5840088 (zugegriffen: 09. Dezember 2022). Vorveröffentlichung.

Jain, Nitisha, Alejandro Sierra-Múnera, Maria Lomaeva, Julius Streit, Simon Thormeyer, Philipp Schmidt und Ralf Krestel. 2022. „Generating Domain-Specific Knowledge Graphs: Challenges with Open Information Extraction.“ In *Proceedings of International Workshop on Knowledge Graph Generation from Text (Text2KG), co-located with the Extended Semantic Web Conference (ESWC 2022)*.

Kejriwal, Mayank. 2019. *Domain-specific knowledge graph construction*. New York: Springer International Publishing.

Mann, Mark, Filip Ilievski, Mohammad Rostami, Aastha und Basel Shbita. 2021. „Open Drug Knowledge Graph.“ In *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022)*. <http://ceur-ws.org/Vol-2873/paper10.pdf> (zugegriffen: 01. August 2022).

Perak, Benedikt. 2020. "Modeling Semantic Relations from a Dependency-Based Graph: A Corpus-Based Network Analysis of Croatian Parliamentary Debates." In *Graph Technologies in the Humanities - Proceedings 2020*. <https://ceur-ws.org/Vol-3110/paper9.pdf> (zugegriffen: 09. Dezember 2022).

Schröder, Markus, Christian Jilek und Andreas Dengel. 2021. "Mapping Spreadsheets to RDF: Supporting Excel in RML." In *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2021) co-located with 19th Extended Semantic Web Conference (ESWC 2021)*.

Wu, Tianxing, Haofen Wang, Cheng Li, Guilin Qi, Xing Niu, Meng Wang, Lin Li und Chaomin Shi. 2020. "Knowledge graph construction from multiple online encyclopedias." In *World Wide Web* 23 (5): 2671-2698.